2016-09

A responsible de-identification of the Real Data Corpus: building a framework for PII management

An, Johanna

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

A RESPONSIBLE DE-IDENTIFICATION OF THE REAL DATA CORPUS: BUILDING A FRAMEWORK FOR PII MANAGEMENT

by

Johanna An

September 2016

Thesis Co-Advisors: Michael R. McCarrin
Dorothy E. Denning

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2016	Master's Thesis 3-13-2016 to 9-10-2016

4. TITLE AND SUBTITLE

A RESPONSIBLE DE-IDENTIFICATION OF THE REAL DATA CORPUS: BUILDING A FRAMEWORK FOR PII MANAGEMENT

5. FUNDING NUMBERS

6. AUTHOR(S)

Johanna An

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: NPS.2016.0040-IR-EP7-A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

De-identification methods have helped government organizations provide the public with useful information—promoting transparency and accountability while also protecting the individual privacy of the data subjects. However, due to the recent massive increase in data collection and improved methods of analysis, de-identification has become a more difficult task. This work outlines challenges and discusses procedures for making a potentially sensitive data set available to extramural researchers and institutions without significant risk to human subject privacy. We provide a detailed explanation of personally identifiable information to help us understand what forms of personally identifiable information can cause the most harm. Furthermore, we discuss the legality and ethics behind working with personally identifiable information to illustrate the importance of protecting privacy. We then offer a taxonomy of threats, vulnerabilities, and impacts and describe how these determine risk. Based on this taxonomy, we develop a framework to assess risk on the Real Data Corpus, a collection of forensic disk images containing personally identifiable information. In addition, we analyze de-identification methods such as pseudonymization and anonymization, and consider re-identification risks. Finally, we apply our framework and methodology to a real-world scenario to determine the risk of data disclosure to an extramural researcher.

14. SUBJECT TERMS

De-identification, risk management and assessment, Real Data Corpus, digital forensics, big data, personally identifiable information

15. NUMBER OF PAGES 113

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

A RESPONSIBLE DE-IDENTIFICATION OF THE REAL DATA CORPUS: BUILDING A FRAMEWORK FOR PII MANAGEMENT

Johanna An
Civilian, National Science Foundation Scholarship for Service Recipient
B.A., University of Washington, 2006

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
September 2016

Approved by: Michael R. McCarrin
Thesis Co-Advisor

Dorothy E. Denning
Thesis Co-Advisor

Peter J. Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

De-identification methods have helped government organizations provide the public with useful information—promoting transparency and accountability while also protecting the individual privacy of the data subjects. However, due to the recent massive increase in data collection and improved methods of analysis, de-identification has become a more difficult task. This work outlines challenges and discusses procedures for making a potentially sensitive data set available to extramural researchers and institutions without significant risk to human subject privacy. We provide a detailed explanation of personally identifiable information to help us understand what forms of personally identifiable information can cause the most harm. Furthermore, we discuss the legality and ethics behind working with personally identifiable information to illustrate the importance of protecting privacy. We then offer a taxonomy of threats, vulnerabilities, and impacts and describe how these determine risk. Based on this taxonomy, we develop a framework to assess risk on the Real Data Corpus, a collection of forensic disk images containing personally identifiable information. In addition, we analyze de-identification methods such as pseudonymization and anonymization, and consider re-identification risks. Finally, we apply our framework and methodology to a real-world scenario to determine the risk of data disclosure to an extramural researcher.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AFF	Advanced Forensics Format
CnA	Certification and Accreditation
CD	compact disc
CD-ROM	compact disc (CD) - read only memory
CFAA	Computer Fraud and Abuse Act
CFR	Code of Federal Regulations
CFPB	Consumer Financial Protection Bureau
CFR	Code of Federal Regulations
CIA-triad	confidentiality, integrity, and availability triad
CIL	Confidentiality Impact Level
CIO	Chief Information Officer
COPPA	Children's Online Privacy Protection Act
CS	computer science
CUI	controlled unclassified information
DEEP	Digital Evaluation and Exploitation
DFRWS	Digital Forensic Research Workshop
DIACAP	DOD Information Assurance Certification and Accreditation Process
DOD	Department of Defense
DON	Department of Navy

DUA	data use agreement
ECPA	Electronic Communications Privacy Act
EU	European Union
FCRA	Fair Credit Reporting Act
FDA	Federal Drug Administration
FERPA	Family Educational Rights and Privacy Act
FIPPs	Fair Information Practice Principles
FIPS	Federal Information Processing Standards
FISMA	Federal Information Security Management Act
FOIA	Freedom of Information Act
FOUO	for official use only
FTC	Federal Trade Commission
GLBA	Gramm-Leach-Bliley Act
GPS	Global Positioning System
GSM	Global System for Mobiles
HHS	U.S. Department of Health and Human Services
HIPAA	Health Insurance Portability Accountability Act
HITECH	Health Information Technology Economic and Clinical Health Act
IA	information assurance
IACIS	International Association of Computer Investigative Specialists
IHE	Integrating the Health Enterprise
IR	internal report

IRB	Institutional Review Board
IS	information system
ISO	International Organization for Standardization
IT	information technology
JPEG	Joint Photographic Experts Group
LES	law enforcement sensitive
MIST	MITRE Identification Scrubber Toolkit
NIJ	National Institute of Justice
NIST	National Institute of Standards and Technology
NPS	Naval Postgraduate School
NRA	National Research Act
NSPII	non-sensitive personally identifiable information
NTFS	New Technology File System
NUS	Non-U.S.
OISF	Ohio Information Security Forum
OS	operating system
PDA	problematic data action
PE	portable executable
PHI	personal health information
PI	personal information
PII	personally identifiable information
PPDM	privacy preserving data mining

PPDP privacy preserving data publishing

PRMF privacy risk management framework

RCE reverse engineering code

RDC Real Data Corpus

PMD predictability, manageability, and disassociability

RMF risk management framework

SBU sensitive but unclassified

SC security category

SD secure digital

SDL statistical disclosure limit

SECNAV Secretary of the Navy

SIM subscriber identity module

SP Special Publication

SPII Sensitive Personally Identifiable Information

SSN Social Security Number

TS technical specification

TTP tactics, techniques, and procedures

UCLA University of California Los Angeles

UCNI unclassified controlled nuclear information

URL uniform resource locator

USB universal serial bus

USC United States Code

Acknowledgments

First, I would like to express my heartfelt gratitude to Michael McCarrin. His dedication to and energy for the computer science program is inspiring, and he deserves to be acknowledged. I would also like to extend a special thank you to Dr. Dorothy Denning. It was an incredible privilege to have discussions with her. Dr. Denning's work in cyberterrorism, information security, and policy played a huge part in me wanting to pursue the SFS program at NPS.

I thank the National Science Foundation for this chance to study computer science, and I am incredibly blessed and proud to live in a country where such opportunities exist. I thank many of my instructors, especially Richard Cote, Victor Garza, the good-humored and caring Prof. Man-Tak Shing, and the truly awesome Chris Eagle.

I was extremely lucky to be a part of an amazing cohort. Despite our rigorous academic schedule, their familiar faces and supportive presence enriched my life and education. "Military Hogwarts for Hackers" was tough but, even with stress, Jennifer, Casi, and I were able to make each other laugh. Ladies, your friendship and support have been invaluable to me. Henry, thanks for the bro talks and the food highs and lows. Carol and Little Dee Dee, thanks for taking care of me.

My family are my fortress of solitude where I regain my strength. I thank my furry roommate Maddy, for being a living, breathing teddy bear. I am the proud sibling of two valedictorians, Jin and Jennifer, who are still my role models. To my hero Alan, thanks for helping me be the woman I want to be. You are already perfect. I thank my beautiful and strong mother, who prays for me every day. And most important of all, I thank my 아빠. He is gone. I know he is proud of me, but I am also very proud of him. I am grateful for his hard work, sacrifices, and love, which still defines me. Thanks for helping me get up 아빠!

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1: INTRODUCTION

1.1 Introduction

Digital forensic tools and methodologies can help organizations understand criminal or other adversarial activity through analysis of digital media. Attribution is a fundamental prerequisite of our justice system; it maintains accountability by leading to punishment of those responsible for adverse effects and deterring others from similar actions. Understanding how an adverse effect occurred gives an organization the opportunity to protect itself and can offer insight into potential vulnerabilities.

Research in digital forensics supports these crucial capabilities in light of an ever growing array of challenges facing forensic analysts, ranging from increasing complexity of systems to increase in the volume of casework and total data requiring analysis. However, the digital forensic researcher faces a separate set of equally formidable challenges. As the quantity of available data increases, digital forensic tools and methodologies have become increasingly complex, challenging computer science (CS) researchers to come up with tools and processes to protect both organizations and individuals. The risks involved with information systems have increased with the advent of data center-scale systems and data sets that grow rapidly in terms of complexity, variety, and size. Legacy computational methods are outdated, and researchers now need new methods that address big data.

Adversaries present threats to several different types of CS researchers. While information security professionals are tasked with protecting personally identifiable information (PII) of data subjects in information systems, adversaries often exploit the PII of human data subjects that reside in information systems for profit, creating an almost symbiotic relationship where the pursuit of one leads to the function of another. That relationship grows constantly in terms of complexity, as adversaries come up with new and better ways to access and exploit PII. PII presents huge problems for other CS researchers who are constrained by ethical concerns for data subject safety. Collecting and working with PII comes with its own inherent risk to individual privacy. Many digital forensic researchers are not interested

in the identity of data subjects, but the mere fact that PII is embedded within data sets precludes them from using or sharing of data sets, a limitation that can impede efforts to demonstrate repeatable results.

For many decades, in order to protect against exploitation of information by adversaries, various de-identification methods have provided human data subjects with anonymity, especially in the areas of health and government. It would not be beneficial to simply keep research data absolutely private, were it even possible, as public or third-party disclosures serve the scientific research community, allowing researchers to share results and test the accuracy and efficacy of methods. Disclosures can also provide a public service; for example, information from the U.S. Census, allows policy makers to check the state of its citizens and determines correct congressional representation of constituents. Results from clinical trial research regulated by the Federal Drug Administration (FDA), provides the public with health and safety information. However, due to what Professor Latanya Sweeney of Harvard's Data Privacy Lab calls the "data-rich network," reverse engineering of de-identification methods can now re-identify what was once thought to be anonymized information [14]. According to the Belmont report, re-identification poses a big ethical problem for those in the research community, who are held to a high standard, when protecting the interests and welfare of human subjects of research [15].

With respect to the challenges presented above, the objective of our thesis is to define the risks associated with the Naval Postgraduate School (NPS)'s Real Data Corpus (RDC) and de-identified disclosure. The NPS Digital Evaluation and Exploitation (DEEP) lab maintains a data set consisting of 65TB of disk images. NPS purchased the RDC data set using secondary storage devices in secondhand markets outside the U.S. The RDC data set contains PII and, due to the risk of identification, the data set is restricted from public access. Identifying and removing PII in drive images remains an unsolved problem due to the complexity of the content and also the heterogeneity of file types contained within drive images. Additionally, the risk of re-identification poses another challenge to data subject confidentiality. To illustrate the difficulty, de-identification of audio or video files, or proprietary storage formats used by executables like hard drives, differs significantly from text-based formats. Even with text-based formats, de-identification techniques alone may not completely eliminate the identity of the data subject due to ever-improving correlation capabilities that leverage big data.

Our research takes a broad approach, establishing a basic risk management process, known as a risk management framework (RMF), which we hope will provide a baseline understanding for researchers that grows over time. In addition, we present a sample risk management process that focuses on the impact of human subject privacy and may help define organizational objectives, research goals, and effective security control measures. The thesis's risk management process may make PII exposure highly improbable, give a certain level of protection to RDC subjects, and also advance the state-of-the-art in digital forensics by providing researchers access to a rich collection of real data.

We aim to enhance state-of-the-art digital forensics research by allowing greater access to real test data while minimizing the risk to the privacy of the data subjects. Since access to the RDC would allow advancements in digital forensic research, it is crucial that researchers find ways to de-identify PII effectively to keep the identity of data subjects confidential, while simultaneously providing availability so researchers can benefit from the data sets.

1.2 Motivation

The benefit of working with a data set obtained from real (rather than simulated) human data subjects is that it gives a high level of real-world information, according to Garfinkel et al. in their article "Bringing Science to Digital Forensics with Standardized Forensic Corpora" [16]. However, the PII contained in the RDC makes sharing of it difficult and requires Institutional Review Board (IRB) approval, which can be time consuming and administratively burdensome. De-identification may provide non-Department of Defense (DOD) researchers with faster, more reproducible results but not without DOD risk of re-identification and loss of confidentiality of its data subjects [16]. Before endeavoring to de-identify algorithmically run results on the RDC, the risks need to be assessed and perhaps measured in some way to minimize the chance of harm. Researchers can conduct risk assessments to determine what risks are present and what may be considered acceptable. This thesis's sample risk management process, as well as the baseline understanding of PII implications and complexities it offers, present a start to one possible solution to PII and the RDC.

Researchers in computer science are not usually clinical doctors nor experts on human or civil rights law. However, the controls they implement and data management practices can

affect people and cause harm. Our research seeks to define those potential harms and risks so that a computer scientist can anticipate and therefore determine the necessary security controls to mitigate such threats or vulnerabilities. The results of our research may help a data provider to determine early on if a request by an outside researcher is feasible or too resource intensive to be delivered. With a baseline understanding of potential real-world effects, digital forensic researchers can complete further work with de-identification and re-identification research using this RMF. By setting up the foundational policy and procedures, this thesis endeavors to eliminate much of the guesswork for researchers and reduce potential for unintentional harm for individuals and organizations.

1.3 Research Questions

Our research explores an alternative to sharing the RDC, which achieves the benefits without the need to release full RDC content to external researchers. To implement this approach, NPS needs a set of criteria for researchers who wish to run analytical algorithms on RDC disk images. If the algorithms adhere to the criteria set out, NPS will create a process of running these algorithms against the RDC and de-identifying the output before returning the results to researchers. The hypothesis is that, if the criteria are sufficiently restrictive (meaning that all output is required to take the form of structured text), we may reduce the risk of PII exposure. Specific questions our thesis seeks to address follow.

- How can we allow extramural researchers access to the RDC and institutions without significant risk to human subject privacy?
- Can we successfully de-identify PII output generated by vetted algorithms provided by external researchers and safely disclose the results?
- At what point do results lose their utility when too much PII is removed?
- What are the risks, and what is considered acceptable risk of disclosure?
- Due to the heterogeneity of data, can we effectively build a criteria for algorithms and how restrictive must the criteria be to protect human subject confidentiality?

1.4 Scope

The scope of this thesis investigates ways in which NPS can share the RDC data with other academic institutions with minimal risk to human subject confidentiality. Initially,

our research focused on de-identification of results derived from programs run on the RDC. Although we endeavored to experiment with a few de-identification scenarios, thorough background reading of PII classification and de-identification processes left us searching for a more comprehensive approach. For example, re-identification attacks break the confidentiality of de-identified PII, rendering the process useless. Although specific experimental scenarios on this topic do provide insights, "extension neglect" [17], or scoping a problem so narrowly by a researcher, that they neglect the complexity or relationships it has with the bigger problem (i.e., big data), can leave anonymized data subjects vulnerable. Understanding that disk images contain a variety of information types, file formats, and data modalities, the diverse requests made by researchers makes de-identification of PII not only difficult, but potentially ineffective. Since de-identification has such a wide-ranging potential impact for both research and privacy we limited the scope of this thesis to establishing a solid foundation of understanding legal, ethical, and other ramifications, and presenting the results of one study.

To help facilitate sharing of RDC, we define many legal, ethical, and regulatory provisions set by the DOD, Department of Navy (DON), or any other applicable authoritative body. With a solid understanding of organizational requirements and context, we formulate criteria for algorithms and security controls that would allow the RDC to be tested by outside researchers and then perform de-identification on test run results. Our intent is to build the beginnings of a cybersecurity and privacy risk framework that will allow NPS to share RDC data in a way that reduces the risk of harming RDC data subjects.

1.5 Significant Findings and Contributions

Our research contributes through the following means. The research:

- Provides a basic framework to responsibly release data sets that contain personal information (PI)
- Frames organizational objectives and other laws and regulations tied with the RDC
- Develops taxonomy of data types and access levels
- Identifies threats, vulnerabilities, impacts, and security controls of the RDC
- Suggests tools for de-identification in other data modalities.
- Provides a list of safe practices when applying various de-identification methods

- Analyzes risk on one RDC scenario
- Establishes a procedure for how PI is processed and recommends continuous monitoring or logging of various PI overtime to improve our understanding of linkage attacks and re-identification.

1.6 Relevance and Contribution to the DOD

The approach and methods in our research can be utilized by any organization that carries the fiduciary responsibility of managing PII and the privacy of their data subjects. The U.S. DOD is a federal government agency and the largest employer in the world, employing 1.3 million active duty service members, 742,000 civilian personnel, 826,000 National Guard members and Reservists, and a fluctuating number of contractors; additionally, the DOD supports two million military retirees and their families [18]. Aside from managing such a large number of personnel and contractors, the DOD's domain of supervision encompasses all military branches, national intelligence services, research and development support centers, educational institutions, and the military health system [18]. With the tremendous responsibility of managing PII on such scale and variety, the DOD also has a mission to provide a level of transparency to its citizens. According to the DOD Principles of Information, the DOD has a full commitment to "make available timely and accurate information so that the public, the Congress, and the news media may assess and understand the facts about national security and defense strategy" [19]. The DOD Principles of Information also highlight things that could potentially threaten the U.S. or violate the privacy of its employees and citizens [19].

As recently as 2012, the DOD made reductions to cease the pervasive use of Social Security Numbers due to increases in identity theft [20]. Our research establishes a foundation of understanding for DOD researchers to apply when carrying out all aspects of their mission regarding PII and data sharing.

1.7 Thesis Structure

Our thesis structure is comprised of eight chapters. Following this introduction, the chapters are structured as follows.

- Chapter 2 gives background and terminology regarding digital forensics, big data, and PII. The chapter goes on to define risk, de-identification, and methods.
- Chapter 3 discusses the legal considerations that relate to data distribution when that data contains PII, also specifically addressing the RDC. It also discusses laws, regulations, and standard bodies.
- Chapter 4 examines related work and how processes like de-identification practices are changing due to emerging threats including re-identification.
- Chapter 5 presents a risk taxonomy for the RDC and discusses the framework for assessing risk.
- Chapter 6 applies the framework to a real-world scenario and makes a determination for disclosure.
- Chapter 7 summarizes our progress towards establishing a baseline for researchers and recommends potential future work.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
BACKGROUND AND TERMINOLOGY

The de-identification of data sets affects multiple areas of study, well beyond the realm of computer science. Chapter 2 defines relevant concepts and conveys contextual information so as to establish a baseline of understanding. We focus first on illuminating concepts within the digital forensics field, as well as contextualizing how our research relates to those concepts, then we define PII and examine various PII data types. Finally, we on define risk, de-identification, and influential factors that aid in the retrieval and removal of PII data from information systems.

2.1 Digital Forensics

Originally applied to law enforcement, digital forensics stems from the broader field of Forensic Science. Mark Pollitt [21] asserts that the term digital forensics did not exist prior to 1985; however, due to the growth of personal computers, digital forensics came to be because the hobbyists in law enforcement at the time saw value in computers as aids to investigations. The International Association of Computer Investigative Specialists (IACIS) was formed in 1989 [21]. The National Institute of Justice (NIJ) defines Forensic Science as "the application of sciences ... to matters of law" [22]. Thus, since its infancy, digital forensics has maintained an evidence-based ethos, endeavoring "to identify, collect, examine, and analyze data while preserving the integrity of information and maintaining a strict chain of custody" to avoid compromising evidence [23]. The study of digital forensics not only encompasses the examination of data, but also places important emphasis on how data is collected. Today, the utility of digital forensic techniques goes beyond the scope of law enforcement. Many organizations apply digital forensic methods to collect and analyze data from various sources in areas such as incident response, asset recovery, and operational problem solving. The National Institute of Standards and Technology (NIST) Special Publication (SP) 800-86 *Guide to Integrating Forensic Techniques into Incident Response* emphasizes that "practically every organization needs to have the capability to perform digital forensics" or "would have difficulty determining what events," such as exposure of protected and sensitive information, "have occurred within its systems and networks" [23].

Like everything digital, digital forensics has evolved rapidly and substantially since its inception in the 1980s, and as any science, is subject to high standards. Digital forensic methods follow scientific standards that place significance on quality of data, not merely the quantity of data processed. Because the results of forensic methods provide substantiation for and are admissible in legal proceedings, the integrity of data and correctness of methods must be proven, and the chain of custody be transparent and well-documented. The 1993 U.S. Supreme Court decision, *Daubert v. Merrell Dow Pharmaceuticals Inc.*, established the "Daubert Standard," which became the basis for the standards for digital forensic methods by the International Organization for Standardization (ISO) and NIST [24]. The standard discusses the legal criteria that constitutes a scientific technique as reliable. These characteristics are empirically tested and peer reviewed, and they require disclosure of potential error and control standards, and acceptance by the scientific community [24]. ISO 5725 and NIST's *General Test Methodology for Computer Forensic Tools* specifies that, in order for a method to be validated, it must state its purpose, and it must go through extensive examination using empirical evidence to assure accuracy, or the trueness[1] and precision[2] of its results [26], [27]. Therefore, procedures for a valid test method ensure that "test results must be repeatable and reproducible" [28].

While digital forensics remains subject to excellent legal standards, it evolves so quickly that the laws that cover and relate to it quickly become antiquated. Conventional digital forensic practices may have adequately addressed legal and civil investigations for evidence recovery in the past; however, this has changed due to the pervasive use of data and digital devices used every day and in every facet of our lives. The landscape of data has changed such that the "scale of data [that] must be analyzed is vast, the variety of data types is enormous ... and the forensic investigator today must make sense of any data that might be found on any device anywhere on the planet" [29]. With such diverse and so much data, it is terrifically complicated to apply laws correctly and keep them up to date. Further legal implications regarding data are covered in Chapter 3.

[1]ISO 3534-1 defines trueness as "the closeness of agreement between the average value obtained from a large series of test results and an accepted reference value" [25].

[2]ISO 3534-1 defines precision as "the closeness of agreement between independent tests results obtained under stipulated conditions" [25].

2.2 Data Science and Big Data

"Big data," a phrase that typically characterizes a problem-solving approach that uses massively parallel analytical methods on terabytes or petabytes of data, often stored in data centers, and may draw on techniques developed in the field of data science field, refers also to the condition produced by this approach, namely of data sets growing rapidly in terms of complexity, variety, and size. Issues with big data impact all areas of the sciences, especially those where the validation of methods rely on real and accurate data sets. Many organizations struggle with data's propensity to be "too big, too fast, or too hard for existing tools to process" [30], or what is more commonly known as Gartner's three Vs: volume, velocity, and variety. Introduced in 2001 and defined below, the three Vs concept is an effort to describe the components that comprise and complexify big data [30] (IBM also added another vector, veracity) [31]

- **Volume** relates to the rapid growth of data being generated, used, and stored. Essentially expanding the scale of data, the growth in disk storage capacity with Kryder's Law[3] has grown from compact disc (CD) - read only memory (CD-ROM) devices, and universal serial bus (USB) flash drives, to millions of data centers and mass migrations toward cloud storage services [32]. Mass consumption of multimedia, increase of quality and size of files, and 2.5 quintillion bytes of data generated every day only adds to the volume.
- **Velocity** refers to the speed of data transmission and how fast it can be processed. Is data transferred through a network in batches, in real-time, or streaming? Is the analysis and processing of data immediate, or does it require storage? The speed of information and analysis has grown rapidly; for example, financial transactions can now be made and verified instantly [31].
- **Variety** refers to the heterogeneity data types. These various data types are primarily categorized as either structured or unstructured (see Section 2.9.1). We discuss the taxonomy of data in greater detail in Section 5.1. The trend, however, is that the variety of data types increases constantly while the majority of data remains unstructured [31].
- **Veracity** refers to the accuracy or integrity of information within a data set. The variety and enormous volume of data makes quality control difficult. Traditional

[3] areal storage density

> methods are inadequate to process unstructured data, and, since the majority of data generated is unstructured, the accuracy of unstructured data sets comes into question.

Digital forensics must overcome problems stemming from all four factors. The digital forensic community has been tackling the challenges of big data and slowly making strides, a concept addressed in Chapter 4.

However, research in forensics faces an additional problem: much of the data of interest contains PII. As conventional digital forensic methods become obsolete or struggle to keep up [29], researchers want larger and larger data sets to develop new, scalable approaches. Unfortunately, large data sets containing user PII in heterogeneous unstructured file formats raise potentially far-ranging legal and privacy concerns. A recent Berkley data science article raises the following question. Can current technology securely and responsibly share large data sets that contain PII [33]? In order for digital forensics to responsibly share large data sets, analysts need current standards to assess potential risk. In the article, researcher Garfinkel makes the point that analysts need to "do more than search and present," and he writes that they also need to "develop new scientific techniques in data analysis, sense-making, machine learning and related fields" [16]. In order to develop tools capable of meeting the needs of law enforcement agencies, government organizations, and private corporations, researchers need access to data sets that reflect both big data and real world conditions.

2.3 PII, PI, and Identifying Information

Although variations in terminology exist (mostly due to the sectoral[4] nature of U.S. privacy laws and regulation), the general term PII connotes information that identifies a specific individual. NIST SP 800-122 defines PII as follows:

> Personally identifiable information is any information about an individual maintained by an agency, including *(i)* any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date,

[4]When referring to laws, the U.S.'s approach to privacy is sectoral, meaning that laws and regulations only apply to certain sectors and are very specific, whereas, for example, the European Union (EU)'s approach to privacy law is more comprehensive and encompasses all industries [34].

and place of birth, mother's maiden name, or biometric records; and *(ii)* any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information [8].

In NIST internal report (IR) 8053, Garfinkel cites inconsistencies with the usage of PII by various groups and states that "personal information is used to denote information from individuals, and identifying information is used to denote information that identifies individuals" [11]. In this thesis, we consider personally identifiable information (PII) and identifying information to have the same meaning. Figure 2.1 shows the relationship between different categories of information relevant to our research: information, personal information, identifying information, and sensitive identifying information. Identifying information or PII is a subset of PI which also means that all PII is PI, but not all PI is PII, or, in other words, not all PI gives a troublemaker the ability to identify someone completely.

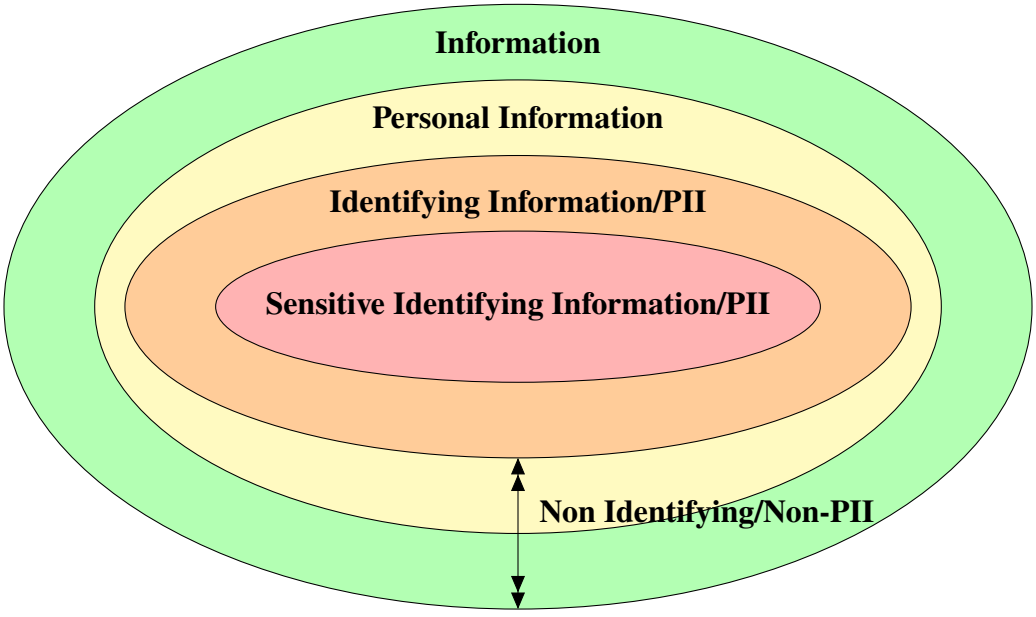

Figure 2.1. "Venn Diagram Depicting Set Relationships of Sensitive-PII and Information Categorization: SPII $\subseteq$ PII $\subseteq$ PI $\subseteq$ Information." Source: [1].

2.3.1 Types of Identifiers

What is an identifier? The ISO defines an identifier as a categorical variable, whose value holds information that may establish identity [11]. Identifiers refer to a "natural person"[5] or persons, also known as data subject(s) [36]. So, identifiers refer to natural persons, and there are several categories of identifiers. The categories of identifiers used in a data set depend on the privacy model, regulations, and the volume and variety of data processed. The process of de-identification for privacy management refers to several direct and indirect (or quasi-) identifiers, including: biometric identifiers, personal health information (PHI) identifiers, financial identifiers, persistent identifiers, and many more [11].

Direct Identifiers

NIST defines *direct identifiers* as *directly identifying data*, meaning that the variable data contained is unique to one specific individual and can, therefore, explicitly identify that individual without other relational or linked information [11]. Examples of direct identifiers include: names, social security numbers, and email addresses. Most policies and regulations stipulate that direct identifiers be removed or redacted from, or at least anonymized within, any documents or other data sets before disclosure.

Indirect Identifiers or Quasi-Identifiers

As opposed to direct identifiers, indirect identifiers (also known as non-unique identifiers or quasi-identifiers) do not alone contain the information to "identify a specific individual" [11]. However, also according to NIST, when "aggregated and linked with other information," indirect identifiers can be used to identify specific data subjects [11]. Some examples of indirect identifiers include ZIP codes, sex, race, and birthdays. If the information contained in an indirect identifier is specific enough, a deduction could be made from a population, thus the identity of a person could be inferred with very few aggregated indirect identifiers [11]. Depending on policy, indirect or quasi-identifiers may not be removed from publicly disclosed material. Garfinkel highlights further trickiness when it comes to quasi-identifiers: they are not as easily identifiable, removing them can harm the utility of the data set, and quasi-identifiers can cause re-identification risk once someone has already gone to

[5]The term natural person is used in ISO technical specification (TS) 25237:2008 and other standards regarding PII, when referring to identifiers, personal data, et cetera. The ISO also makes the distinction between natural persons and legal persons. Legally, a natural person refers to a human being, while legal persons are entities, such as organizations or companies, that have some duties and legal obligations but may not carry human rights [35].

the trouble of removing them; therefore, he assesses that data controllers need to weigh the risks between potential re-identification and utility when dealing with quasi-identifiers [11].

2.3.2 Sensitive and Non-Sensitive PII/Identifying Information

Finding the identity of a specific person online using PI is relatively effortless and common. A vast majority of technology users have digital footprints and established online identities. Also, even people that choose not to establish themselves online can be subject to having their directly identifying data (their names, addresses, and names of family members, for example) publicly available. Some states also publish voter registration information, so choosing to register to vote may carry the risk of having your home address and political party preference published [11]. However, not all displays of personal information online, whether the person chooses them or not, carry the same risk. To a malicious attacker, some types of PII are more valuable than others and can be monetized, often with detrimental effects to the individual. Depending on the hacker's intention, gains other than financial may be sought. Therefore, establishing a qualitative metric of sensitivity regarding publishing data allows researchers to evaluate risk and measure the impact of disclosure [8].

Sensitive Personally Identifiable Information (SPII)

The DOD defines Sensitive Personally Identifiable Information (SPII) as any information about a person which would, if lost, stolen, or compromised, present a significant risk or could cause harm to an individual, so that non-sensitive personally identifiable information (NSPII)[6], as opposed to SPII, is perceived to be "minimal or non-existent" [6]. For reference, DOD-defined examples of PII and whether they are sensitive or non-sensitive are illustrated in Table 2.1. The directive continues to point out that not all PII exposure may cause harm and that NSPII falls under this category. NSPII also identifies an individual but such information may already be public. Details on what constitutes minimal risk and harm are discussed in Sections 2.5.3 and 2.5.4.

[6]Non-Sensitive PII is also known as *Internal Government Operations* or *Business PII* [6].

Table 2.1. DOD List of Sensitive and Non-Sensitive PII. Source: [6], [7].

Sensitive PII	**Non-Sensitive PII**
Names - *official or others*	Location of an office
Citizenship, legal status	Business email address
Gender, race-ethnicity, sexual orientation	Zip Code
Birth date, place of birth	Business telephone
Home, personal cell phone numbers	Business cards
Personal home, email, mailing addresses	Published work or projects
Religious affiliations or preference	Employment history in resume
Security clearance	Badge numbers
Mother's maiden and middle names	Schools attended and graduated
Government ID numbers - *driver's license, full or partial social security, passport, etc.*	Memberships and donation info
Marriage and family - *spouse, marital status, children, emergency contact*	
Health records - *medical, biometric, disability, insurance*	
Financial information - *credit cards, account numbers, types of accounts*	
Law enforcement information	
Educational records - *student info, grades, transcripts, class schedules, billing*[7]	

2.4 Confidentiality, Integrity, and Availability

In addition to potential harm to individuals, we face potential large-scale harm from cybersecurity attacks, and the U.S. government has increased relevant laws and systems according to the CIA security objectives. In response to the growing number of cybersecurity attacks on U.S. information infrastructures, the Federal Information Security Management Act (FISMA) was enacted in 2002 to build an information security framework and standardize federal information systems [37]. An information system (IS) is comprised of

[7] A complete listing can be found within the Family Educational Rights and Privacy Act (FERPA) statute.

information resources that manage all information processes of an agency and provide operational support to help facilitate its mission [37]. FISMA defines information security as the "protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability" [9]. The confidentiality, integrity, and availability triad (CIA-triad) of security objectives provide three key foundational principles that guide information security decisions [37]. According to FISMA, confidentiality refers to "preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information" [37]. Integrity refers to "guarding against improper information and modification or destruction, and includes ensuring information non-repudiation and authenticity" [37]. Finally, availability refers to "ensuring timely and reliable access to and use of information" [37].

Tasked by FISMA to produce the Federal Information Processing Standards (FIPS) 199, NIST helped federal agencies secure and evaluate their information and ISs. FIPS 199 directs agencies to first take stock of their information and then to define and place that information into categories called information types [9]. For example, where an IS may be processing credit card information, the information type would be financial. In addition, FIPS 199 guidelines incorporated security categories where information types can be measured against the "potential impact" if there was a loss in CIA-triad security objectives [9]. This concept is further addressed in Section 2.5.3. The potential impacts show what potential consequences the agency would face to its assets, processes, and people. Equation 2.1 illustrates how FIPS 199 performs a security categorization using a security category (SC) information type [9].

$$\textbf{SC}\ \text{information type} = [(\textbf{confidentiality}, \textit{impact}), (\textbf{integrity}, \textit{impact}), (\textbf{availability}, \textit{impact})] \tag{2.1}$$

2.4.1 PII Confidentiality Impact Level

Given different levels of potential impact a security breach might cause, NIST SP 800-122 established guidelines called the *PII Confidentiality Impact Levels (CILs)* on how to cat-

egorize PII based on potential risk and impact due to loss of confidentiality. When PII "sensitivity" level is high, if compromised the loss of "confidentiality" would yield catastrophic harm in the CIL classification table. NIST uses the legal definition of confidentiality, meaning "preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information" [38]. Table 2.2 illustrates how, when PII is evaluated under PII CILs guidelines, it is placed into one of three categories: *low*, *moderate*, or *high* [8].

Table 2.2. Personally Identifiable Information Confidentiality Impact Level Category Designations and Definitions. Source: [8], [9].

PII Confidentiality Impact Levels		
Low	Limited Adverse Effect	Only minor harm to human subjects. Effectiveness of function is noticeably reduced. Small degradation of capability
Moderate	Serious Adverse Effect	Significant harm. No death or serious life injury. Significant degradation and capability.
High	Severe or Catastrophic Adverse Effect	Severe catastrophic harm. Loss of life and serious life threatening injuries. Major financial loss. Not capable to perform.

As researchers process vast amounts of data, direct or quasi identifiers may be hard to determine. Even if a specific identifier is identified, understanding how a breach or accidental disclosure could harm an individual may not be obvious. Referring back to NIST SP800-122 applying these factors in Table 2.3 will help define and categorize PII.

Table 2.3. Contributing Factors that Determine PII CILs. Where d-ID is a Direct Identifier and q-ID Means Quasi Identifier. Source: [8].

Factors	**Definition**	**Impact**
Identifiability	How easily PI can uniquely identify an individual.	d-ID has greater impact than q-ID
Quantity of PII	Number of human data subjects affected or the size of the compromised data set. Depending on PII sensitivity and context of use, quantity may be factored less.	Impact on 1,000 data subjects greater than 10
Data Field Sensitivity	Evaluate the sensitivity of PII data fields (i.e., database). Also evaluate linking factor between multiple q-ID data fields to obtain identity.	single or combo

Context	Purpose for PII and how it is used - FIPPS principled.	PI from classified info is higher risk than unclassified
Obligations for Protect Confidentiality	PII disclosure dictated by specific laws, policy, or code of ethics.	Release of SSN greater than release of grocery store ID number
Access to Location of PII	Figuring out impact by how accessible information is and probability for breach due to location.	PII stored in cloud increases impact than directly controlled PII

2.5 Risk and Organizational Risk Assessment Considerations

To further understand the complexities inherent in data and privacy, we define and contextualize the notion of risk, as what constitutes risk (let alone the notion of unacceptable or acceptable risk) necessarily varies given an individual's or organization's objectives. The ISO defines risk as the "effect of uncertainty on [an entity's] objectives," where an effect can present either positive opportunities or negative threats and vulnerabilities or a combination of both [39]. The ISO's explanation of risk may be viewed as panoptic, but it is common sense to note that objectives necessarily vary depending on the particular entity's mission, resources, and levels of strategy. Organizations can have different risk attitudes [40], mindsets regarding how to handle and evaluate risk. Although most organizations seek to benefit from opportunities, some organizations measure risk by performing risk evaluations geared toward measuring potential loss and, therefore, their risk attitude is designed to scrutinize components that could jeopardize a system. To use a very simple example, if the main priority of organization *A* is to protect sensitive information and preserve confidentiality, then *A*'s risk attitude would be considered risk averse [41]. Since the impact of loss in confidentiality is high, organization *A* may decide not to take the risk. If the objectives of another organization, *B*, required quick and reliable access to information (availability), *B* might be more willing to consider trade-offs. The risk attitude of *B* would be to incur higher levels of risk while seeking benefits from opportunities, known as increasing risk appetite [40]. Ultimately, for any organization, risk assessment is likely to boil down to

weighing potential benefits against potential consequences, as well as the severity of impact of those consequences on organizational assets and objectives.

To somewhat mitigate the complexity inherent in defining risk and provide further context, additional definitions follow. Less ambiguous than the ISO's definition, the NIST SP 800-37 defines risk as

> a measure of the extent to which an entity or individual is threatened by a potential circumstance or event, and typically is a function of: *(i)* the adverse impact that would arise if the circumstance or event occurs; and *(ii)* the likelihood of occurrence [42].

Definitions of risk provided by NIST's standards tend to align more with, or give more weight to, security and privacy objectives, which focus on protection from loss of confidentiality. As stated in Section 2.4, the quality of confidentiality focuses on preserving access to a "secret", therefore any PII leak or exposure loss may be irreversible. Residual risk is the remaining level of risk after mitigating factors, such as security controls, safeguards, or countermeasures have been implemented [2]. In the case where risk is completely unmitigated before or without controls, the level of risk is known as inherent [43].

Without context, measuring risk is a nebulous task. Therefore, it is the organization's responsibility, with regards to their circumstances, to frame, assess, respond, and monitor potential risks, a process known as *risk management* [5], shown in Figure 2.2. According to NIST SP 800-30, risk is assessed within a risk management strategy when organizational objectives, responsibilities, assets, and participants are fully framed [2]. The purpose of a risk assessment, then, is to identify and measure risk so as to make less risky decisions. Assessing risk means evaluating and making prudent assumptions about threats, vulnerabilities, impacts, and the likelihood of harm throughout all levels of an organization [2].

2.5.1 Risk Management Framework (RMF)

The DOD, like any organization, requires a solid process, or framework, to assess and manage risk, commonly known as an RMF. Before 2015, the DOD Information Assurance Certification and Accreditation Process (DIACAP) was the standard information assurance body that evaluated all DOD ISs for cybersecurity risks and privacy protections [44].

DIACAP's Certification and Accreditation (CnA) process authorized IS operations, conducted security assessments, and provided regulatory compliance [44]. Although providing risk management, DIACAP's view on risk was static and lacked the continual monitoring that would make the RMF adaptable to changes in the system [42]. While some risk factors remain constant, like a bank always risks a robbery and, therefore, should always account for that risk by implementing security controls to mitigate potential loss, other risks fluctuate. Also, while the greed that motivates bank robberies may be a constant, the tactics, techniques, and procedures (TTP) utilized by adversaries change over time [42]. As organizations increase their technological and operational capabilities, adversaries are likely to find weaknesses in the new systems; therefore, risk is inherently dynamic [42].

In 2015, in response to the need to have a more adaptable RMF that considered risk as dynamic, the DOD adopted NIST SP 800-37 [42]. NIST SP 800-37 provides an RMF for federal organizations that offers an adaptable system life cycle approach where constant monitoring throughout all levels of an the enterprise provides faster response to operational changes or risk [42]. It is important to note that frameworks encompass the management of a whole organization, starting with a broad approach to managing a three-tiered system where Tier One is the organization or governance structure, Tier Two the logistical/operational mission layer, and Tier Three is the information systems and control level [42]. The DOD's RMF is a six step process, laid out below.

- Step 1: Categorize information systems by its information types. Some information systems contain more sensitive PI than others (i.e., health versus social) [42].
- Step 2: Select the foundational security controls that will help protect the information and allow it to operate initially. Depending on the needs of the information system, some security controls promote more access control while others focus on availability [42].
- Step 3: Implement the security controls and document their need and operation [42].
- Step 4: Assess if security controls that were implemented are functioning properly and protecting information systems [42]. Table 2.2 shows the steps to risk assessment.
- Step 5: Authorize the determination made after a risk assessment. Consider if the implemented remediations from the assessment phases are "acceptable" [42].
- Step 6: Monitor, a crucial component to RMF. Observe whether security controls are working effectively or are growing outdated due to new threats. Monitoring

allows security controls to evolve and adapt to changes and factors in risk's dynamic properties [42].

The DOD's RMF focuses on identifying and mitigating potential risks within the confidentiality, integrity, and availability triad (CIA-triad) security model. The RMF covers loss of confidentiality issues, while also focusing primarily on cybersecurity threats and vulnerabilities in regard to unauthorized access [2]. The DOD's RMF is outlined in Figure 2.2.

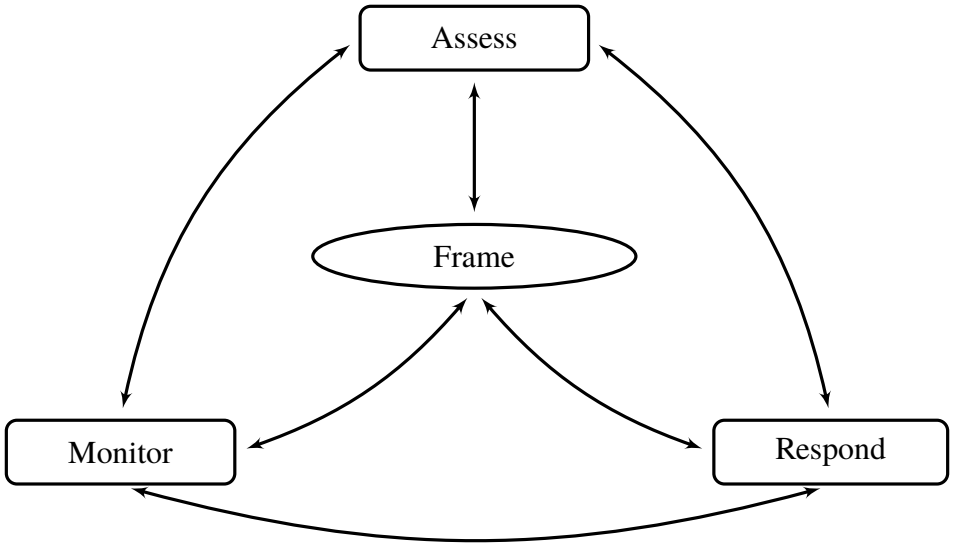

Figure 2.2. NIST Risk Management Process Depicting the Four Steps of Risk Assessment. Source: [2].

2.5.2 Privacy Risk Management Framework (PRMF)

Since organizational and individual risk management differ, the DOD also implemented NIST IR 8062, a risk management methodology specifically designed to evaluate privacy risks for individuals [3]. The privacy risk management framework (PRMF) is influenced by the RMF but differs in scope, drawing a distinction between cybersecurity and privacy risk management, where adverse impacts are caused by an IS's operations including how information is processed rather than by a breach in confidentiality [3]. NIST IR 8062 notes that factors that contribute to privacy risk fall outside what is typically considered a threat or vulnerability area [3]. To address these differences, the PRMF establishes privacy risk

engineering objectives and a privacy risk model that includes four phases [3] as shown in Figure 2.3.

The four phases and six processes of the PRMF can be understood as follows. The first phase is to frame risk, first by business objectives and then by organization. Business objectives help an entity serve its purpose by setting up a beginning to end risk management strategy while organizational privacy forms the legal and privacy oriented structure to operate [3]. Then, an organization evaluates both its functional and privacy risks, using risk assessments [3]. Risk and risk components are identified, including threats, vulnerabilities, harm or impact to subjects, and the probabilities of each [3]. The calculations resulting from the assessment are used for risk determination [3]. The third phase, where privacy controls are as designed, focuses on how to safeguard or reduce privacy risk, which can be technical controls or defined by Fair Information Practice Principles (FIPPs) [3]. Risk response considers and compares compare risk with organizational objectives, such as risk tolerance, and responds with appropriate action [2]. The fourth phase, monitoring change, happens after the controls are implemented and keeps watch of how personal information is managed [3]. Since the PRMF focuses exclusively on privacy risk factors, NIST IR 8062 advises that an organization should execute a RMF strategy concurrently with PRMF to defend against unauthorized access [3]. The PRMF redefines privacy risk in reference to what NIST IR 8062 calls *data actions* which are "IS operations that process personal information. [Processing] can include, but is not limited to, the collection, retention, logging, generation, transformation, disclosure, transfer, and disposal of personal information" [3].

Privacy Risk Model Equation

Generally, when organizations conduct quantitative risk assessments, risk is calculated by: *(i)* the probability that an event may occur, and *(ii)* the event's potential impact [3]. In terms of PII and risk management, the NIST IR 8062 draft on PRMF provides the formula for privacy risk as [3]:

$$\text{Privacy Risk} = \sum_{\text{All data actions}} \text{likelihood of (PDA)} \times \text{impact of (PDA)} \tag{2.2}$$

The PRMF privacy equation measures risk through problematic data actions (PDAs). A

PDA is "a data action that causes an adverse effect or problem, for individuals" [3]. For example, if an organization collects more PII than what is needed for them to operate, it has the potential to cause even more and unnecessary harm to the individual if there is a breach [3].

Privacy Engineering Objectives

NIST IR 8062 also provides a privacy security model. The privacy security model in the DOD's PRMF has a similar focus as information security's CIA-triad: it focuses on three privacy-preserving ISs, predictability, manageability, and disassociability (PMD) [3]. *Predictability*, "is the enabling of reliable assumptions by individuals, owners, and operators about personal information and its processing by an information system" [3]. *Manageability*, "is providing the capability for granular administration of personal information including alteration, deletion, and selective disclosure" [3]. *Disassociability* "is enabling the processing of personal information or events without association to individuals or devices beyond the operational requirements of the system" [3]. The PMD model is a privacy preserving model that uses the principles set by FIPPs.

PRMF is currently being drafted in NISTIR 8062 and although standards in privacy preserving models are in their nascency, privacy risk is becoming a bigger concept now due to the pervasive dissemination and collection of PI by the Internet of Things, companies, and legitimate entities [3]. Additionally, the loss of control over an individual's information or even attributes allows linkages to occur and makes re-identification attacks possible as described in Chapter 4.

2.5.3 Risk Equation

The risk equation (see Equation 2.3), is useful in qualitative risk assessments because it helps rank and categorize. Reducing risk into smaller categorical components helps us identify characteristics and aids in risk model and policy development.

Although not solely used for confidentiality or privacy protection, the risk equation is a common construct in information assurance when referring to principles of risk management [45].

$$Risk = \frac{Threats \times Vulnerabilities \times Impact}{Security\ Controls} \tag{2.3}$$

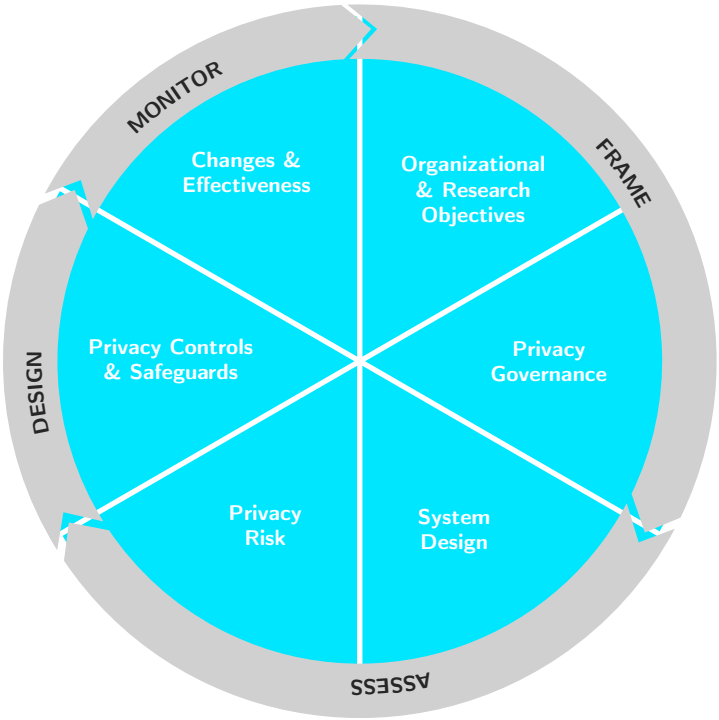

Figure 2.3. NIST Privacy Risk Management Framework (PRMF) Diagram: Built on top of the RMF, this diagram shows six distinct cyclical processes that organizations can utilize to responsibly secure and protect the privacy of ISs and data subjects. Source: [3].

Threats

A threat is comprised of a threat source(s) and threat event(s) [2]. NIST SP 800-30 explains that a threat originates from a *threat source(s)*, a source that exhibits "*(i)* the intent and method targeted at the exploitation of a vulnerability or *(ii)* a situation and method that may accidentally exploit a vulnerability" [2]. If the threat source(s) are highly probable and effective, this triggers the threat event, which could be a single or set of events, actions, or circumstances [2]. The definition and potential consequences of a threat event are contained within the definition of threat provided by NIST SP 800-30:

> any circumstance or event with the potential to adversely impact organizational operations and assets, individuals, other organizations, or the Nation through

> an IS via unauthorized access, destruction, disclosure, or modification of information, and/or denial of service [2].

NIST SP 800-30 also provides a list of threat source(s). These examples may be categories that require more subdivision if necessary. They are:

- purposeful/hostile cyber or physical attacks
- environmental disruptions
- human errors of omission and commission/machine errors
- structural failures of organization-controlled resources which can include: hardware, software, environmental controls, and failed security controls
- man-made disasters, accidents, and failures beyond the control of an organization [2]

When threat events and sources are identified, an organization is able to contrive hypothetical *threat scenarios* where, especially in the case of adversaries, brainstorming of TTP occur. Identifying potential TTP provides threat characteristics which help organizations construct threat taxonomies [2].

Vulnerabilities

As mentioned within the NIST definition of threat, a *vulnerability* is a predisposed weakness in a system that could be exploited inadvertently or with intent, by a threat source or combination of sources [2]. Vulnerabilities can include a weakness in a specific program or a combination of flaws that exist throughout all levels of a system or organization. Some common examples include social engineering, where an employee is manipulated into granting privileged access, or poor system design or project execution where security controls were never implemented [2]. In information assurance (IA) training, the vulnerability component in the risk equation (see Equation 2.3) is the point at which an organization's defense intersects with an adversary or threat [45].

Impact

NIST SP 800-122 defines *impact* as follows.

> The magnitude of harm that can be expected to result from the consequences of unauthorized disclosure of information, unauthorized modification of information, unauthorized destruction of information, or loss of information or

information system availability [8].

In regards to the risk equation, we measure impact in regards to the level of harm. Harm is inflicted on either the individual or an organization. Organizational harm, as opposed to individual *harm* refers to "any adverse effects that would be experienced by an individual whose PII was the subject of a loss of confidentiality, as well as any adverse effects experienced by the organization that maintains PII. Harm to an individual includes any negative or unwanted effects (i.e., that may be socially, physically, or financially damaging)" [8].

NIST SP 800-122 provides the following as examples of impact placed upon an individual: "blackmail, identity theft, physical harm, discrimination, humiliation, or emotional distress" [8].

Security Controls and Safeguards

Organizations implement security controls and safeguards in order to reduce the impact of harm. Their implementation can reduce the likelihood of an unwanted event. Security controls can come in various forms including policy, operational, or technical controls [3]. Security controls can include a statistical algorithm designed to hide identity within a dataset (i.e., differential privacy) or anything that masks or performs de-identification on an identifiable person [11]. De-identification is also a security control, which provides confidentiality to the data subject while also making that data set available to researchers.

2.5.4 Minimal Risk

The DOD and ethical codes on human subject research use the term minimal risk as a determinant for public disclosure [6]. The DOD defines minimal risk as a situation where

> the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests [46].[8]

[8]DOD Instruction 3216.02 makes extended clarification that (section 219.102 *(i)* of [U.S. Code of Federal Regulations (CFR)]) "shall not be interpreted to include the inherent risks certain categories of human subjects

Many risk assessments use minimal risk, also known as acceptable risk, as a threshold for what information can be disclosed to the public.

2.6 Pseudonymization

One way to protect privacy in a data set containing PII is through pseudonymization. Pseudonymization is a form of masking that obscures or replaces the original direct identifiers with artificial data or some symbolic place holder (i.e., hashes, numbers, letters, codes) [8]. Between NIST SP800-122 and NIST IR 8053, there is some discrepancy as to whether pseudonyms are considered de-identification. Garfinkel in NIST SP 800-188 specifically addresses pseudonyms as not de-identified [48]. Pseudonymization is useful because it generally removes the identity of the real data subject while still preserving some of the links or relationships in the data set [11]. Pseudonyms should be used only if the identity of a data subject needs to be preserved due to the objectives of the data controller, who may use encryption to keep identity confidential, and then later de-crypt when identification is necessary. It is important to address that the Common Rule does not view "coded" pseudonyms as anonymous [3].

Pseudonymization is considered by Garfinkel as a trade-off between protecting privacy and preserving data set utility but can be viewed as less secure or more prone to re-identification attacks [11].

2.7 De-identification

Another form of protecting privacy in a data set containing PII is de-identification. Similar to the sectoral variations and differing usages of the term PII, "de-identification" does not have a single authoritative definition. As Garfinkel states in NISTIR 8053, de-identification is often confused or used interchangeably with terms like anonymization, pseudonymization, and sanitization [11]. The ISO standard definition of de-identification is "any process of removing the association between the set of identifying data and the data subject" [36]. De-identification is achieved through the removal or changing of PII in a data set, so data

face in their everyday life. For example, the risks imposed in research involving human subjects focused on a special population should not be evaluated against the inherent risks encountered in their work environment (e.g., emergency responder, pilot, soldier in a combat zone) or having a medical condition (e.g., frequent medical tests or constant pain)" [47].

subjects cannot be identified [8]. Processes described in the following sections all facilitate de-identification but differ in the ways they handle identifying data.

It is important to note that de-identification does not always mean the identity of a data subject cannot be recovered. In fact, the legitimacy of de-identification is being called into question due the emerging threat of re-identification, described in Section 4.1 [49].

2.7.1 Anonymization

A subcategory of de-identification, anonymization is one specific way researchers protect privacy in a data set containing PII. Like de-identification, anonymization of a direct identifier disassociates the data subject from the data set, essentially removing all of the connecting links [48]. In NIST IR 8053, Garfinkel explains that anonymization is irreversible, so, while it protects privacy, if that data is ever needed, it will no longer be available [48].

2.7.2 Sanitization and Redaction

Another way to protect privacy in a data set containing PII is through sanitization or redaction. According to Garfinkel, sanitization is the erasing or the overwriting of information on a hard disk [50]. Depending on the file system, erasing a file may mean the operating system (OS) frees the block but leaves the file data until it is written over by another process [50]. Overwriting not only frees blocks that contain the file data but overwrites them using ASCII NULL or more sophisticated Guttman patterns [50]. Also a form of de-identification, redaction removes or blacks out information [47].

2.8 Health Insurance Portability Accountability Act (HIPAA)

One particular subset of PII, PHI has long been considered sensitive and has more controls around its regulation. The U.S. Department of Health and Human Services (HHS) website defines PHI as information regarding a person's mental or physical health at any period of time or information about health care [51]. For instance, names or Social Security Number (SSN) which are direct identifiers when framed within the context of health records become PHI, while quasi-identifiers are not considered PHI [51]. PHI is specifically covered by the HIPAA Privacy Rule, which are laws that dictate how PHI should be handled by

various entities, known as "covered" [52]. It is important to note that once PHI is de-identified then the privacy rule's restrictions do not apply to that data set and can be shared [10]. However, due to re-identification HHS has stipulated that once de-identified PHI is uncovered, then the privacy rule still protects that data set [10]. HIPAA conducts their de-identification standard using two methods: expert determination and Safe Harbor [51]. HIPAA's methods may provide DOD researchers additional context for understanding how de-identification can happen.

2.8.1 Expert Determination

HIPAA uses expert determination, which provides a statistical method of anonymizing identities. An expert determination to disclose information can only be made by a statistical expert who can confirm that the risk is minimal [51]. Since health professionals often have more background on human resource issues than CS researchers, implementing expert determination within the DOD would require additional personnel.

2.8.2 Safe Harbor

HIPAA's other method, the Safe Harbor method, does not require a statistical expert, and so poses another option for organizations to disclose data sets responsibly [51]. The Safe Harbor method of de-identification happens through the process of eliminating 18 identifiers, listed in Figure 2.4.

2.8.3 Limited Data Set

While HIPAA's Safe Harbor method requires that the data controller perform de-identification on 18 identifiers, a limited data set may not require the removal of quite so many attributes [10]. That limit does not include direct identifiers, however. Limited data sets can be disclosed after removal of 16 identifiers listed in Table 2.6, with the researcher signing a data use agreement (DUA) [10].

The data use agreement must:

- Agree to use a limited data set consistent with reason why it was disclosed
- Name those who are authorized to use the limited data set

Table 2.4. Safe Harbor Privacy Rule for De-Identification Comprised of 18 Identifier Types. Source: [10]

18 Identifiers for De-Identification	
1.	Names.
2.	All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
2a	The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people:
2b	The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
3.	All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4.	Telephone numbers.
5.	Facsimile numbers.
6.	Electronic mail address.
7.	Social security numbers.
8.	Medical record numbers.
9.	Health plan beneficiary numbers.
10.	Account numbers.
11.	Certificate/license numbers.
12.	Vehicle identifiers and serial numbers, including license plated numbers.
13.	Device identifiers and serial numbers.
14.	Web uniform resource locators (URLs).
15.	Internet protocol (IP) address numbers.
16.	Biometric identifiers, including fingerprints and voice prints.
17.	Full-face photographic images and any comparable images.
18.	Other unique identifying number, characteristic, or code, unless permitted by the Privacy Rule for re-identification.

- Recipient needs to give assurances that they will not share it with unauthorized users, have and maintain safeguards, employee monitoring, and not contact or try and reach out to individual [10]

Limited data sets are shared among trusted researchers with the stipulation of signing a

DUA. The removal of too many quasi-identifiers can affect the utility of the data. A DUA enables more data to stay intact, promoting more effective research by keeping information, like age, intact.

Table 2.6. Limited Data Set required to De-identify 16 Identifier Types from National Institute of Health. Source: [10].

16 Identifiers for De-Identification	
1.	Names.
2.	Postal address info, other (town or city), state, and ZIP Code.
3.	Telephone numbers.
4.	Fax numbers.
5.	Electronic mail addresses.
6.	Social security numbers.
7.	Medical record numbers.
8.	Health beneficiary numbers.
9.	Account numbers.
10.	Certificate/license numbers.
11.	Vehicle identifiers and serial numbers, including license plate.
12.	Device identifiers and serial numbers.
13.	Web universal resource locators (URLs).
14.	Internet protocol (IP) address numbers.
15.	Biometric identifiers, including fingerprints and voice prints.
16.	Full-face photographic images and any comparable images.

2.9 Information and PII

2.9.1 Structured and Unstructured Data

When evaluating information systems, computer scientists categorize information as structured or unstructured. For instance, information contained in a relational database is structured, because each field has some value associated, and when a query is made, a result is returned [53]. For unstructured information, a query may not return an answer because there is no column, row, or field to find. Unstructured data like logs, images, and arbitrary length text documents pose a problem because, if a tool cannot identify specific PII, then we cannot de-identify it.

2.9.2 Storage Devices

Secondary storage devices are comprised of non-volatile memory used for long term storage needs such as hard drives and flash drives [54]. A secondary storage drive may be comprised of one or more volumes which house a file system [55]. These file systems contain various files whose structure is determined by the OS and applications that created them [54].

To preserve a secondary storage device, a person creates a disk image, which is a "sector by sector copy of a disk" that can retain all file system info [56]. When a hard drive is imaged, it is typically done using a write-blocker with proprietary or open source imaging software like Advanced Forensics Format (AFF) and EnCase [57]. Many of these forensic tool kits do more than image; some do file system analysis and more.

2.9.3 `bulk_extractor`

One of the tools we will be using for disk image analysis is bulk_extractor. bulk_extractor is a powerful forensics tool that, as its name hints, extracts detailed information from various inputs such as disk images, directories, and files [58]. It can extract information from compressed files as well, which many similar tools cannot [58]. As this tool was developed for law enforcement, the extraction of identifying information was the goal [58]. bulk_extractor is capable finding e-mails, URLs, credit card numbers, Global Positioning System (GPS) coordinates, or simply an entire listing of words contained within [58]. Along with the extraction, it provides a histogram, a count of every instance that a particular piece of information or word was present [58].

There are three phases to the operation of bulk_extractor. The first is called the feature extractor, where the information that is being searched for is extracted and then written to a text file [58]. Next, a histogram is produced, counting every instance of the feature that was found [58]. Lastly, the post-processing phase produces the readable report [58].

It is easy to see how this wealth of extracted information might give ability to identify the owner of the drive. One proven technique is that extracting the e-mails within a system and sorting to find the result with the highest instances typically identifies the owner of the file system [58]. Other typical uses are utilizing the word count analysis to aid in password cracking [58].

2.9.4 MITRE Identification Scrubber Toolkit (MIST)

Free form text is difficult to analyze when compared to structured data. MIST is a de-identification tool that leverages natural language processing to analyze documents [59]. Natural language processing is a broad term referring to the application of computational methods to analyze human language. This tool searches through text, chooses potentially identifying phrases and then de-identifies them [59]. MIST works by first annotating, or finding, potential areas of concern using natural language processing [59]. Once these areas are annotated, its next task is to replace, swapping words or phrases to accomplish de-identification [59]. For example, if the names of individuals on a list needed to be de-identified, the processor would annotate every instance of a name and either replace it with a pseudonym or a generalization such as "[NAME]" [59].

CHAPTER 3:
LEGALITY AND ETHICS

3.1 Privacy

If asked, most people would likely have some notion of what constitutes privacy, though, of course, their notions would differ. Certainly, the notion of privacy has changed over time; while the World Wars' generations tended to consider privacy more of a human right, at least in the United States, the current generations willingly post private details. This shift necessitates a revamped understanding of both ethical and legal privacy considerations, one which acknowledges an individual's right to be left alone and the desire for freedom and discretion to disclose personal information without intrusion, while accounting for the fact that what may be considered private to one person may not be for another.

Chapter 2 posits that the definition of privacy is elusive because it is a dynamic social construct and, therefore, requires context to establish a baseline of understanding to consider the ramifications of de-identifying PII within data. In the digital age, users frequently divulge personal information that should be kept private. Personal information is often divulged to public and private organizations simply in order to receive services. For convenience, our computers also store a lot of user information. When considering how and whether to de-identify PII from data, it is necessary to consider what PII even looks like on a hard drive. In what format is it stored, and how much or in what location is it saved? While those fundamental considerations are necessary for CS engineers' consideration and planning, first we must look at the potential ramifications and how we have been and currently deal with and legislate PII concerns. In Section 3.2, we discuss PII with regards to privacy law, ethics, and the RDC. The following sections define terminology and elucidate how societal and legal concepts of identity are transposed to data types.

3.2 Legal and Ethical Concerns for Research on Data Containing PII

Legal and ethical issues hinder the sharing of data sets containing PII, and regulations vary by region, state, and area of business (assuming we confine our discussion to the U.S.). Part of the difficulty can be attributed to the fact that, as the Privacy, Data Protection and Cybersecurity Law Review observes, there is “no single omnibus federal privacy law in the U.S." nor a "designated central data protection authority” that ensures and protects a citizen’s PII as a fundamental right [60]. The U.S. implements sectoral, or area-specific, privacy laws which “regulate only a specific context of information use” in particular areas, both in public and private sectors [61]. Privacy and data protection in areas such as finance[9], health-care[10], electronic communications[11], educational records[12], and privacy of minors[13] have regulatory frameworks to guide people’s actions [60]. Historically, risk has always been unavoidable and is taken into account when dealing with the management of sensitive PII in such areas [60]. Although privacy protections exist, the system of privacy regulation and governance is scattered throughout various departments on federal and state levels, each defining their own rules and regulations for PII [61], which does not exactly aid in consistency.

The U.S. judicial system and individuals with grievances also play a part in shaping privacy law [60]. Private litigation holds organizations accountable and deters those industries that collect, store, and use PII from unfair, negligent, and deceptive business practices [60]. Attorney Alan Charles Raul states that the “U.S. privacy system is flexible, relying more on *post hoc* government enforcement and private litigation;” additionally, he adds that “the U.S. system does not apply a precautionary principle to protect privacy, but rather allows injured parties to take legal action” [60].

Despite providing an overview of area-specific regulations, Raul’s assessment of U.S.

[9]Laws associated with PII in financial sector Gramm-Leach-Bliley Act (GLBA) laws, Federal Trade Commission (FTC) Act, Consumer Financial Protection Bureau (CFPB), Fair Credit Reporting Act (FCRA)

[10]Laws associated with PII defined in the healthcare sector HIPAA, Health Information Technology Economic and Clinical Health Act (HITECH)

[11]Laws that define PII in electronic communications are Electronic Communications Privacy Act (ECPA), Computer Fraud and Abuse Act (CFAA)

[12]Education laws that dictate PII usage FERPA

[13]Children’s Online Privacy Protection Act (COPPA)

privacy laws do very little to help the digital research community develop state-of-the-art tools and methods, leaving researchers with more quandaries than answers.

A dynamic judicial process with checks and balances may have advantages, but, from a researcher's perspective, with limited resources and murky definitions, the possibility of a lawsuit naturally stifles progress. Much of Garfinkel's work on the RDC dealt with trying to navigate through this legal and ethical landscape. Garfinkel found what information privacy lawyers like Paul Ohm also observed: the current sectoral approach to privacy laws left out entire industries from definitive privacy regulations [49]. The inherent ambiguity is problematic for researchers who count on the scientific methodology of repeatable and reproducible results to validate methods and build on foundational work. If data sets cannot be shared due to the repercussions of disclosed PII of data subjects, and researchers have no means to mitigate or understand how PII may cause harm, scientists are hindered from performing experiments on real data sets. Without means to validate methods on real data, digital forensic tools may work on contrived scenarios but fail when put to operational use, which reduces their practical value.

3.3 The Belmont Report

The Belmont Report was established by the National Commission for the Protection of Human Subjects on Biomedical and Behavioral Research in 1979 and is the ethical framework adopted by the human subject research community [15]. The framework consists of three fundamental principles: "respect for persons, beneficence, and justice" [15]. As stated, much of U.S. laws on privacy are sectoral especially in regards to digital privacy where individual protections are not clearly defined [62]. Due to the Belmont report, those who conduct research using human subjects are ethically bound and follow HHS's policy known as the Common Rule. Researchers have the responsibility to maintain and not adversely affect the welfare of their human subjects (beneficence), therefore, an assessment of risk should be conducted to the best of their ability to determine if the research is justified [15]. It is the Belmont report that introduces the term minimal risk as defined in Section 2.5.4 and echoed throughout DOD instructions.

3.4 The Real Data Corpus

An endeavor started by Simson Garfinkel, the RDC was created with the hope of offering a standardized data set for the digital forensic research community [16]. The scale and diversity of data is growing, and, because data is user-driven and generated, the RDC provides researchers with a representative sampling of drives characteristically similar to those found in the real world.[14] Comprised of 3,098 disk images, the RDC is a rich collection of raw data extracted from various devices such as hard drives, flash memory images that include USBs, secure digital (SD) cards, memory sticks, CDs, digital camera memory images, and Global System for Mobiles (GSM) subscriber identity module (SIM) chip images, all formatted and stored as EnCase Evidence File denoted by extension identifier ".E01". The disk images contain a variety of data in different file formats from common document files in various languages, graphic, or video file formats like Joint Photographic Experts Group (JPEG) and .mp4, and also binary executables. Many of these RDC files contain PII of individual users and their disclosure may cause harm.

Because NPS is a research institution, privacy rules concerning the RDC go beyond the controls implemented by the typical academic institution. NPS purchased all devices in the RDC from the secondary market, outside the U.S., and those devices contain data collected from non-U.S. persons and very likely contain various types of PII [16]. The U.S. Supreme Court decision of *California v. Greenwood 486 U.S. 35*, in 1988, held that items discarded or sold in secondary markets do not have reasonable expectation of privacy [63]. Therefore, sharing the RDC would not be illegal, regardless of whether data contained private user PII and regardless of ethical concerns. However, NPS is a research institution. The RDC was established for the purpose of scientific research and education, so NPS researchers are bound to ethical codes of conduct for human subject research.

Additionally, NPS is a U.S. Navy school, one of a handful of academic institutions under DOD purview, and, therefore, subject to additional controls, including review and permission by the IRB before any human subjects research can be conducted. Since the RDC is federally funded and contains authentic data and "identifiable private information" from real living people, any research conducted on the RDC must abide by the National Research

[14] Garfinkel notes that images bought from secondary markets may have more instances of drive sanitization, corruption, disk failure or reformatting, which researchers should account for because it ties into the motivations of why a used item was resold [16].

Act (NRA) of 1974 and title 45 CFR part 46, the Common Rule, which considers user-generated data to be human subject research [64]. That includes the research contained in this thesis and any research conducted on the RDC by NPS. Since the RDC was pre-collected NPS research may fall under "existing" data, defined in 45 CFR 46.101(b)(4), where an exemption can be made if "the information be recorded by the investigator... in such a manner that the subjects cannot be identified, directly or through identifiers linked to subjects" [65]. Although this thesis investigates responsible methods of de-identification on the RDC to reserve data subject anonymity, 45 CFR 46.101 requires that, after research and use of such data sets, all information considered identifiable (including the original source) will have to be destroyed, which does not serve the research objectives of NPS with reliability and reproducibility [66]. Most Common Rule investigations are also required to seek informed consent of the data subject [66]. NPS research is not able to do this, due to the purchasing of hard drives via the secondary market. Our research was given exemption from 45 CFR 117(c)(1) [65]. This thesis seeks to de-identify all PII produced from results, and our risk tolerance is aimed at not releasing anything above minimal risk. The U.S. HHS grants IRBs approval authority on any experimental research involved with human subjects and holds them responsible for upholding HHS ethical standards and policies on human subject protection [67]. The RDC is maintained by NPS with approval of access determined by its IRB. NPS also falls under DON jurisdiction, which has considered the collection as controlled unclassified information (CUI) as well as for official use only (FOUO); those item are discussed in more detail in Section 3.5.

Any human subjects research at NPS is under further controls and standards. As a component of the DOD, NPS is subject to DOD Instruction 3216.02, which concerns conducting human subject research [47]. Specifically, DOD Instruction 3216.02 defines a human subject as an "individual about whom an investigator conducting research obtains data through intervention or interaction with the individual or obtains identifiable private information" [47]. As a Navy institution, NPS must also abide by Secretary of the Navy (SECNAV) Instruction 3900.39D, Human Research Protection Program, which states that the "rights, welfare, interests, privacy, confidentiality, and safety of human subjects shall be held paramount at all times and all research projects shall be conducted in a manner that avoids all unnecessary physical or mental discomfort, and economic, social, or cultural harm" [46]. It is essential for an organization to define its objectives when assessing risk, especially when building

the foundations of an RMF. Research and education are, necessarily, not the only DOD objectives. However, the DOD recognizes the RDC's value for research and refers to the Common Rule standards within their policies.

3.5 Controlled Unclassified Information

In addition to controls via ethical and DOD human subjects research laws, the RDC is considered CUI. CUI is unclassified nonetheless restricted in distribution because, in some cases, according to DOD 5200.01, some unclassified data may"require application of access and distribution controls and protective measures for a variety of reasons" [68]. CUI was established by Executive Order 13556, Controlled Unclassified Information, which replaces the sensitive but unclassified (SBU) classification. Executive Order 13556 was implemented to create uniformity amongst classification categories across all departments in the executive branch, specifically regarding unclassified information [69]. CUI now includes the control and protection of the following types of unclassified information: FOUO, law enforcement sensitive (LES), DOD unclassified controlled nuclear information (UCNI), and limited distribution; these types of CUI all fall under DOD CUI [68] and are subject to the rules and policies of the DOD Information Security Program [68]. In general, some characteristics that require information be considered CUI include: needing to be reviewed and approved before public release, potentially export-controlled, and potentially considered to have permanent value as a record [68].

Containing sensitive information, and warranting special handling under DOD guidelines, the RDC is considered CUI and, more specifically, FOUO. FOUO, a type of CUI, is designated so because it can potentially cause harm due to the possibility of violating certain Freedom of Information Act (FOIA) protections [68]. Specifically, the FOIA protects information as FOUO that "the release of which would reasonably be expected to constitute a clearly unwarranted invasion of the personal privacy of individuals" [68] and that concern is why NPS considers the RDC as FOUO. Some of the controls implemented to protect FOUO are: ensuring valid reasons for access, marking appropriately, and taking certain security precautions [68].

3.6 Fair Information Practice Principles (FIPPs)

With all that protection, it is a wonder that NPS considers any outside research requests for the RDC, or, for that matter, that NPS students and faculty can access the RDC. It is important to note that, although the data subjects in the RDC hold no U.S. citizenship, the RDC still needs to be considered under FIPPs. Experts in the field of de-identification and PII management refer to FIPPs as a foundational schema to design privacy-preserving information systems [3]. FIPPs guidelines, produced by the U.S. Federal Trade Commission, are principles that address fair PII collection and protection practices [70]. The first core principle is transparency, where people should be aware how, who, and where their PII is being collected [70]. The second principle is choice, which "gives individuals a choice as to how their information will be used" [70]. The third principle is information review and correction, which allows people to check and access their personal information to correct inaccuracies [70]. The fourth principle, is information protection, where organizations have the legal responsibility to protect the integrity of people's PII [70]. The fifth and last principle is accountability; organizations must comply with FIPPs [70]. Garfinkel states, in NIST IR 8053, that de-identification does not warrant notification to the data subject, although this may be contestable [11].

3.7 Organizational Level Security Controls

Because they may still be subject to financial or other liabilities for wrongfully disclosing PII, many research institutions implement a data use agreement (DUA) as additional protection. A DUA is a contractual agreement usually with a third party that outlines special terms of disclosure regarding the de-identified data set by the data provider [11]. NIST IR 8053 gives a few examples where a DUA states that the data provider would bar a data requester from re-identification and blanketed sharing to others, and would accept all liabilities, including privacy violations due to mismanagement of the data set [11]. DUA should be used in the following situations.

- When it is considered a Limited Data Set [11]
- If the re-identification of risk seems probable

Many research institutions, universities, and government agencies are utilizing DUA as insurance beyond IRB approval to prevent third party from re-identification.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
RELATED WORK

Chapter 4 highlights some of the related work regarding protecting PII within information systems as well as regarding the current state of de-identification methods and regulations. Many researchers address vulnerabilities in de-identification and information disclosure structure. In addition, de-identification researchers are currently studying re-identification to enhance the way that organizations handle information. Chapter 4 first looks at re-identification, then privacy perserving models, and, finally, the possibility of synthetic data sets, which, inherently, would contain no real PII.

4.1 Re-identification

Simple de-identification practices used to be adequate. Practices in de-identification before what Harvard Professor Sweeney calls today's "data rich network" proved mostly effective at providing the public with useful data sets without invading the privacy of individuals [71]. Both big data and re-identification now pose serious threats to conventional de-identification practices. For example, organizations in the past believed that the removal of direct identifiers would make data subjects anonymous in a dataset. In Sweeney's *Simple Demographics Often Identify People Uniquely*, she demonstrates how, even with the removal of direct or "explicit" identifiers, quasi-identifiers in combination with *background information* [11] have the possibility of uniquely identifying an individual [71]. What troubled proponents of de-identification more was Sweeney's example of how knowledge of only a few attributes (she demonstrates using five digit ZIP codes, sex, and birth date) could identify 87% individuals of the U.S. population specifically [71]. Sweeney's research only utilized data sets that were freely available to the public, or available with a nominal fee [71]. In NIST IR 8053, Garfinkel describes Sweeney's scenario: she was aware that a high profile politician was ill at a hospital [11]. She obtained a list of de-identified patients who were discharged combined with a list of local city voter registration information and was able to identify the politician [71].

4.1.1 Linkage Attacks

Linkage attacks occur because the combination of quasi-identifiers or attributes provide uniqueness by forming links to re-identify what was previously de-identified data [71]. For linkage attacks to work effectively, adversaries require two data sets that contain the same data subject(s) [11]. Linkage attacks can uniquely identify one person if the same quasi-identifiers from those data sets have one match [71]. Multiple matches of data subjects in those datasets means that an adversary could associate each with a probability, or be able to narrow down potential data subjects, a potential demonstrated in Figure 4.1 [11]. Since de-identification of direct identifiers alone was not sufficient to protect data

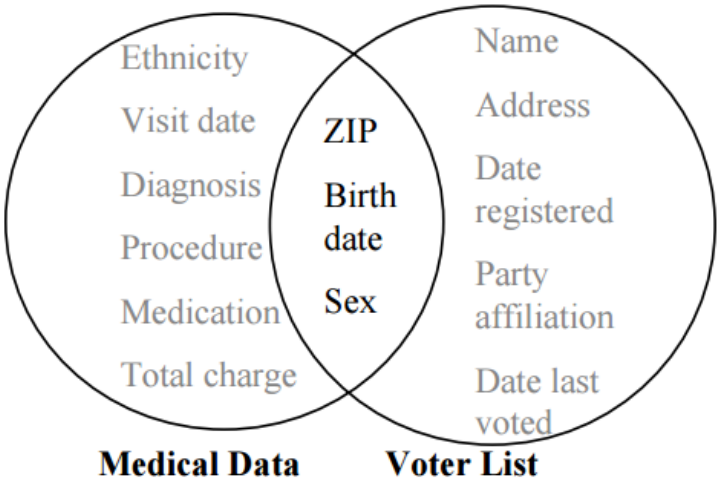

Figure 4.1. Sweeney's Linkage Attack Using Medical Data and Voter List. Source: [4].

subject confidentiality, quasi-identifiers would then need de-identification [11]. However, the removal of several quasi-identifiers could significantly reduce the utility of a data set [11]. NIST IR 8053 lists five ways of de-identifying indirect identifiers: suppression, generalization, perturbation, swapping, and sub-sampling [11].

Re-identification risk dependencies are:

- Data modalities and content of original data set
- Type of de-identification technique used by data controller
- Adversary's level of skill

- Adversary's resources
- Availability of other external data sets for links
- The risk of re-identification over time as more attributes are made available [3]

According to NIST SP 800-188, re-identification probability "is the probability that an attacker will be able to use information contained in a de-identified dataset to make inferences about individuals" [48].

4.2 BitCurator Project

Currently, researchers are exploring different methods of safe de-identification. University of North Carolina Associate Professors, Lee and Woods, in their paper, *Automated Redaction of Private and Personal Data in Collections*, explore methods of de-identification, specifically the BitCurator project. The BitCurator project is an endeavor by Information and Library Science researchers to investigate how data collecting institutions can acquire, preserve, and provide access to collections while protecting private information embedded within such collections [57]. Similar to the objective of this thesis, researchers wanted to find a balance between the protection of private information while retaining the ability to access collections [57]. Resembling other organizations who strive for transparency (availability), Information and Library Science researchers saw that PII poses a huge obstacle for institutions like libraries, so they began the BitCurator project which aims to identify and provide curators with automated software and reporting methods so that they can be better stewards of digital information [57]. Lee and Woods identify the PII problem because they feel that information collecting institutions will lose credibility if they cannot properly care for digital content, and this may hurt their ability to acquire digital collections or cause them to face increasing resistance from producers [57]. Their paper defines private and non-private data and goes into detail about identification and redaction of such material using open source forensic tools, in particular bulkextractor, fiwalk, Sleuth Kit, and sdhash [57].

As the tools were originally designed for digital forensics, Lee and Woods observed that these tools would not specifically meet the needs of the Information and Library Scientist. First, the output of forensics tools does not necessarily lend to digital archival needs [57]. The workflows, or methods in which data integrates into the archival systems, would need to be looked at for compatibility [57]. Second, the tools did not adequately answer the

concern of how to make their findings public, allowing portions of the data to be accessible while not revealing PII [57].

Lee and Woods did have success in using the open-source tools discussed to identify their collections. In their scenario, they felt the forensics tools yielded a reasonably de-identified product [57]. Although with remaining instances of private information, they still had to analyze the risk of disclosure.

4.3 Privacy-Preserving Models

In previous section, we looked at de-identification techniques, but there are other, more complex ways of anonymizing data subjects. Information systems that are geared more toward protecting individual privacy and safer de-identification practices are called privacy preserving models [11]. The goal of privacy preserving information systems is to simultaneously provide confidentiality to individual data subjects while also providing available information to the public. Both privacy preserving data mining (PPDM) and privacy preserving data publishing (PPDP) have their advantages and are used for different reasons as discussed in this chapter [11].

4.3.1 Privacy Preserving Data Mining (PPDM)

PPDM achieves anonymity using statistics and aggregation [11].

Statistical Disclosure Limitation (SDL)

A method of privacy preservation, statistical disclosure limit (SDL) makes use of a few different techniques. One technique that SDL uses is generalization, taking specific data and replacing it with a broader term [11]. For example, the height of an individual can be replaced with the term "tall" or "short," or replaced with a height range. Another technique is to swap data within similar types of information [11]. An example of data swapping is to interchange the ages of individuals on a record, which would not pose a problem for the researcher if age was not a research factor. Finally, SDL adds noise to the data to obscure actual information [11].

Differential Privacy

Another method of privacy protection is called differential privacy, which helps quantify de-

identification in privacy protection [11]. It creates anonymity by adding non-deterministic noise to an appropriately sized data set [11]. Using a characteristic called the degree of sameness, identity is lost through aggregation [11]. Similar to the noise generation in SDL, differential privacy hopes to effect enough change in the data to preserve privacy, while minimizing the effects on accuracy.

***k*-anonymity**

K-anonymity, another privacy preserving model, focuses on quasi-identifiers [71]. K-anonymity is achieved by using an equivalence class method. For every combination of quasi-identifier that can be made, there are "k" matching records, making it anonymous. This model is prone to unsorted matching attacks, a complementary release attack, and a temporal attack where k-anonymity loses its ability to provide privacy. This loss of privacy occurs when the equivalence class lacks proper diversity or if the adversary has background knowledge of the system [72]. To mitigate privacy loss, an equivalence class has to be diverse but also has to be distributed appropriately [72].

4.4 Privacy Preserving Data Publishing (PPDP)

Similar to privacy preserving models, privacy preserving data publishing (PPDP) is a de-identification technique where PII is substituted with either partially or fully synthetic data [11]. There are two methods of creating synthetic data, both using the original dataset as a source [48]. One method, which produces partially synthetic data, uses data swapping and generalization similar to SDL, while the other method, which produces fully synthetic data, creates a dataset based on a modeled version of the original [48].

4.5 Data Release Models

Another way to safely share data containing PII is through the use of data release models, similar to the DUA discussed previously. In NIST SP800-188, Garfinkel gives examples of data release models, which are forms of control that limit how data sets are used in order to reduce risk and to prevent re-identification [11]. Many research institutions already conduct information sharing in such ways. Garfinkel's five examples of data release models are described below [48].

The Release and Forget Model

The release and forget model certainly gives no accountability to the data provider. In Ohm's *Broken Promises of Privacy*, he explains how some organizations have very little follow through other than simply running the de-identification process and providing disclosure [49]. Ohm also shows a data controller's failure to address the utility factor in the data set [49]. With the release and forget method, control of PII is completely lost to the provider and, therefore, completely in the hands of the receiver [48].

The Data Use Agreement Model

A legal document signed between data controller and third party. Depending on the agreement, the data controller may be able to restrict any re-identification attempts made by the party or control further sharing of the de-identified dataset [11].

Simulated Data Verification Model

Offering a limited, simulated data set is another data release model. In this model, a data controller provides a simulated de-identified data set, similar to the original, for disclosure purposes only [48]. Outside researchers can run programs and query the disclosed data set, but, for further verification, researchers can request that their tools be run by the data controller on the original source [48]. The data controller is able to run and test the researcher's results against the original dataset, then release the results using SDL [48].

The Enclave Model

The enclave model requires the most from the data provider. In the enclave model, a qualified data controller would accept requests from reputable extramural researchers, run those requests on de-identified data sets, and report the results to the researcher, never having to share the data set itself. [48].

Interactive Query Interface

An interactive query interface model makes a synthetic dataset releasable to the public (or to a limited segment of the public) by using various privacy preserving methods [48]. Differential privacy may be added to datasets that retain original data, adding noise and providing confidentiality to data subjects, or fully synthetic datasets may also be utilized [48].

These are not technical controls but controls implemented on a operational or organizational

level.

4.6 Impact of Data Set Selection

While researchers could simply use synthetic data sets, and sometimes do, synthetic data sets will never exactly match real data sets. Choosing what type of data set to use is key to the integrity of any research. Sometimes data controllers do choose synthetic data; however, working with synthetic data sets also requires careful study and an exceptional skill level for those disclosing their research. "Synthetic and artificial data sets pose a challenge to researchers and the general public. A synthetic data set designed to allow research on hospital accidents nationwide might let researchers draw accurate, generalizable conclusions about the impact of training and doctor's work hours on patient outcomes, but make it mathematically impossible to identify specific patients, doctors or hospitals. Such a data set would be useless for the purpose of accountability or transparency" [73].

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
TAXONOMY AND FRAMEWORK

In previous chapters, this thesis has established a baseline of understanding regarding the complexities surrounding PII in data sets. Chapter 5 discusses how to approach risk assessment problems, presents a taxonomy of factors to help the researcher identify methods to mitigate risk, and elaborates on factors of both risk and impacts of harm included in the taxonomy.

5.1 Taxonomy

Table 5.1 lists the classifications and definitions used to evaluate our privacy risk scenarios. Regarding NPS's RDC, a few factors to consider include: levels of access provided to the recipient, the type of interface provided to NPS, and the type of output given to the recipient.

Table 5.1. Taxonomy of Risk Scenarios

Classification	**Definition**
Levels of Access	
Nothing/No Access	NPS provides no access or sharing of RDC to non-DOD qualified researchers
Access to Structured file formats	NPS provides qualified researchers access to run vetted algorithms on text-based file formats only
Access to Unstructured file formats	NPS provides qualified researchers access to run vetted algorithms on text-based file formats only
Access to Image file formats	NPS provides access to run algorithms on image file formats and text-based file formats from qualified researchers
Access to Audio and Video file formats	NPS provides access to run algorithms on all of the above formats and audio and video file formats
Complete Access- all data	NPS provides access to run algorithms on all data in the RDC

Interface Classification	
Self-written code	NPS produces its own code to run on RDC and provides requested data to qualified extramural researcher.
Known or Commonly Used code	NPS utilizes known or familiar tools, either open source or proprietary programs, to extract data requested by extramural researcher
Run arbitrary code, provide source	NPS runs code provided by qualified extramural researcher and also receives source code
Run arbitrary code, binary only	NPS runs code provided by qualified extramural researcher with no source code and only binary file
Output Classification	
Limited or Fixed strings	The output of algorithmic data is structured and limited in scope and human readable. Predictable and well organized, the data in structured file formats are easier to identify and mine for data. Some examples of structured data are lists of file-names, metadata, and network packets in language formats like .xml, .txt, .html, and ASCII.
Arbitrary Text	Unstructured text-based data that lacks a certain level of organization or format which makes parsing and identifying PII indicators difficult. Examples include: the whole contents of emails(narratives), Word and PDF documents.

Binary File Formats	File formats where the majority of information is stored as binary data. Binary data by itself lacks significance and translating to produce substantive data relies on various applications. Binary file formats move away from human readable formats like ASCII text. The structure of data depends largely on the algorithm that produced the output and can be structured or unstructured in nature. Examples of binary files are compiled files like images, compressed files, media files, and even text files [74].

5.2 Risks Associated with Data Types and Levels of Access

As stated in Section 3.4, the potential of the RDC as a research resource obligates research institutions to protect human subjects and their personal information from harm. In order to protect the confidentiality of RDC subjects, PI must be anonymized. Because efficacy of de-identification depends on file formats, we classified levels of access from the requester by file types.

No Access

Referring to Figure 5.1, giving No Access to a requester is the safest option to protect subject PII. The No Access classification is currently in place, and, while it does guarantee subject PII protection, it does nothing to provide any benefit to digital forensic research.

Structured data

Giving slightly more access, access to structured data, would not be difficult. De-identification of PII is far more effective with predictable structured string and plain-text file formats because tabular or categorical data is already organized and queried. Open source tools with PII identification features use string matching algorithms to identify personal information. Previous works by NPS faculty and known digital forensic tools like bulk-extractor, referenced in section 2.9.3 offer some assurance that sharing RDC data without PII is feasible. Granting users and their algorithms access to structured and well known text

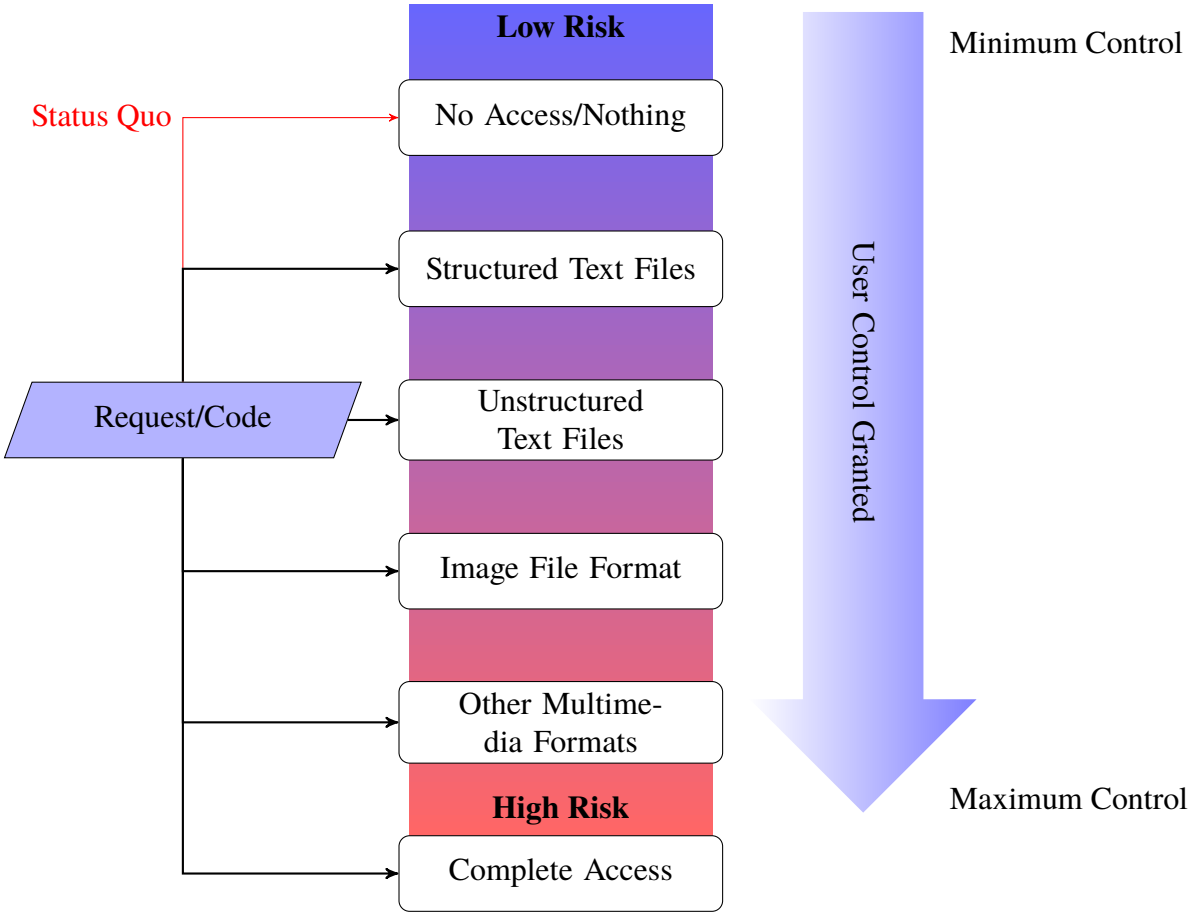

Figure 5.1. Illustration of Risk Increase as Requester/User is Granted Increasing Access to Various RDC Data.

file formats makes it easier for automated and manual checking from the data controllers perspective and makes more effective de-identification possible.

Unstructured data

Unstructured text files like Word documents with long narratives are harder to monitor, and, as Garfinkel stated in NIST SP 800-188, "finding such identifiers and distinguishing them from non-identifiers invariably requires domain specific knowledge" [48]. Although data controllers will de-identify the results produced by a specific algorithm, the ambiguity of quasi-identifiers in the de-identification process makes unstructured text files harder to validate. Part of the process of assessing risk is to categorize our responsibility regarding each information type in our information system. When it comes to analyzing a database

filled with health information, context and structure are predictable. However, a single disc image can contain many different types of PI direct or quasi-identifiers in many different file formats. We would need to know what the requester's algorithm is looking for, and, although levels of access files are not de-identified at this stage, for manual checking, human readable formats provide less uncertainty.

Image or Audio File Formats

Image formats are more difficult to sanitize and require different de-identification techniques depending on what information is requested. Image file objects have more complexity and are rich with various identifiers beyond the usual name or telephone number. Add to that different formats which come compressed or use alternative vector graphics, and allowing image or audio file access becomes even more murky. Some image formats also produce metadata, which, on top of visual PII, produce other PII indicators like geotagging or MakersNote info on Exif files. Creating an automated process to identify PII in these unstructured formats is difficult and an endeavor which exceeds the resources of this thesis because it would require extensive knowledge on such image formats and the binary data produced. However, various multimedia formats, such as image files, may be made accessible to certain requests where the requesters algorithm seeks image file metadata and produces text-based output for easier scrubbing of PII.

Other Multimedia Formats

While image files are considered multimedia, we categorically separated images from video because there are additional complexities which would require *multi-modal de-identification*[15]. Adding another modality increases the number of personal identifiers. Also, if the modality is not well understood, more personal indicators would be vulnerable to insufficient de-identification and ultimately re-identification by possible linking between multiple modes. Therefore, multimedia files pose a significantly higher risk to privacy of subjects, and accessibility should be restricted.

Complete Access

Complete access produces the highest risk because the requester has unlimited access, which would include all files, executable and system files, as well. Heterogeneity of data

[15]Garfinkel defines multimodal de-identification as the combination of biometric identifiers (face, fingerprint), soft biometrics (age, weight), and non-biometric identifiers (hair and dress style) [11]. The combination of these could lead to a full identification of an individual and needs to be de-identified. [11]

and format types overwhelm our ability to adequately find all personal identifiers or possible factors for re-identification. Formats with different encoding and reliance on digital forensic tools that fail to thoroughly de-identify PII pose a high risk for PII leakage or exposure. With certain modalities, especially in the realm of multimedia, identifying personal data requires expertise in other fields such as biometrics and information processing of such data. Machines may also record one event in several different ways (Semantic data). Unfettered access and testing on RDC could make piecing together or linking quasi-identifiers much easier, jeopardizing the identity and personal information of human subjects.

5.2.1 Interface Classification

After a researcher has made a request to access the RDC and communicated clearly their research, objectives, and data needed, the next step is to figure out the logistics of extracting such information from the RDC. As stated in Chapter 1, to reduce the labor of having to de-identify raw data from hard drives, the model proposed would restrict the input and export of data while the de-identification phase would occur on the output of the accepted input algorithm. Our thesis uses an approach that could be likened to the "enclave model," described in Section 4.5. However, we are not running these algorithms on a de-identified database. Therefore, we place emphasis on vetting the input algorithms and performing a security evaluation if a source or binary is given.

Interface classification describes a generalized list of the possible forms the analytical method or program can take and illustrates the risks or benefits associated with a given option.

Using Code Produced/Defined by Naval Postgraduate School

An algorithm written by the provider of information is likely the safest option. The provider of RDC data knows exactly how their program extracts information and has ample ability to test the program in various stages. This avoids any questions about what the program is doing, and the provider would not have to conduct a security evaluation. While potentially the safest option, writing a program for requesters is very time consuming and depends on the skill set of the provider.

Utilizing Known-Code or Open Source/Commercial Software

Another option would be to use familiar commercial software or open source programs like EnCase, Sleuth Kit, Bulk Extractor and fiwalk. These sophisticated forensic tools are commonly used by law enforcement and forensic examiners and are useful software with features that would allow providers options for de-identification. Such tools save the provider from writing their own de-identification program, and such programs have multi-threading features that allow for faster processing. The BitCurator project paper highlights, however, that heavy dependence on such tools does not provide sufficient de-identification because "working with heterogeneous forms of data from many sources" requires a lot of post processing and patchwork, where even sophisticated tools lack compatibility or customization [57].

Open source programs, on the other hand, despite many benefits including being easily accessible, free, transparent, and customizable, also present some vulnerabilities. Since source code is made visible to the public, malicious users may find vulnerabilities in the program and utilize those bugs for PII leaks. Open source tools may have not undergone "meticulous evaluation" that commercial tools have undergone due to industry standards. In addition, contributions made by the public on open source tools may taint the open source distribution with malicious code. Therefore, the use of open source programs needs to be made judiciously by the provider possibly including a security evaluation made on the open source tool before use [75].

Arbitrary Code, with Source Code Provided

The risk of privacy exposure increases when the requester provides their own code. If the requester is utilizing the RDC to do developmental testing on a tool, they may want not only the data but they may also need the performance diagnostics. A security evaluation on the program can be attempted, but this is difficult, and the risk is dependent on the complexity of the code. In this option, the source code is provided, which we may test to ascertain risk and determine if it falls within acceptable risk level. In these types of scenarios, there is a risk that the requester might try and surreptitiously extract RDC PII and hide the data through output. To avoid this, a certain level of confidence must be established through evaluation, testing, and visibility of the program.

From Arbitrary Code, with Binary Only

The types of analytical programs that pose the highest risk to PII leaks are those provided

in only binary file formats. Once an executable program is compiled, it very difficult to conduct a security evaluation because the provider has no design information or visibility on what the code is actually doing. Although reverse engineering approaches exist, there is really no automated or expeditious shortcut that would accurately render source code recovery. Depending on the program's complexity, reverse engineering code (RCE) may be an intense resource and time-consuming endeavor.

Using static or dynamic techniques only help elucidate basic control flow characteristics of a program. The provider could perhaps use a disassembler[16] or decompiler[17] for source code recovery; however, decompilers have no proof of correctness and are considered unreliable [76]. The provider could produce a similar program, essentially reverse engineering the binary via static analysis, utilizing hexeditors, objdumps, or even going with a hybrid or dynamic analysis method of viewing portable executable (PE) headers or running the program through debuggers. This process, however, would be extremely time-consuming given the complexity of the program. Despite reverse engineering methods, even then, there is still a risk that the analysis of the program's behavior might have a PII leak because there is no complete visibility.

5.2.2 Output Classification

Our data sharing model takes algorithms and allows them to run on the RDC to provide their own output. Although a disk image is not an database, on some level, if an algorithm produces a structured output, it makes the data controller's de-identification job a bit easier. This provides an advantage over other information systems that go through the process of complete de-identification before any queries. The length and structure of output are critical to the success of de-identification. Therefore, algorithms that produce structured data, which is limited to a fixed set of strings or typical metadata, is the easiest to work with. Unstructured text-based outputs, such as free form text, and even more problematic binary file formats, reduce the confidence of de-identified results because personal identifiers become harder to classify, match, or even find.

Structured, Categorical, Tabular Data

[16] machine code to assembly code

[17] machine language to some source code, most do not claim take output and feed into decompiler and get same input. Signature profile compiler to create original binary, compiler options.

Structured data can be stored in some field or record and can be easily retrievable. Structured short data is easiest to de-identify because information types are pre-defined categorically which makes it easier to locate PI.

Free Form Text

Text based files with little organization, usually seen with narratives, documents with non categorized data. "Scrubbing" tools like MIST, with natural processing language algorithms, are able to de-identify various identifiers in free form text-based files, especially in areas of health information. In addition, bulk_extractor can identify keywords to bridge the gap between processing unstructured data.

Binary File Formats

All data on a computer is stored in binary, 0's and 1's. The symbolic representation and interpretation by an application of binary data is what renders it either human readable or readable to some other machine or application. In order to identify PI for successful de-identification, readability of data is a necessity.

When working with binary file formats, the risk to PII exposure is substantially reduced if the binary format, with relative ease, can translate data to some human readable form. However, since translation of binary data is application-specific, there are some problems providers face when performing de-identification.

Binary compatibility where various parts of code in a file (data section) can be interpreted the same way but other parts (i.e., file header) may have different information [74].

5.2.3 Possible Threats and Impacts

Tables 5.2, 5.3, and 5.4 list different threats to consider and the potential impacts of disclosure.

5.2.4 Data Categories

Some of the most common data types found in the health field, which can be easily applied to our overall work on de-identifying PII, are listed by the Integrating the Health Enterprise (IHE) information technology (IT) Infrastructure Handbook. Table 5.6 summarizes the common data categories related to PII, discussing examples of each and methods to mitigate.

Table 5.2. Re-identification Scenarios. Source: [11].

Re-identification Threat Scenarios	
Prosecutor scenario	Attacker knows that a specific person is in the dataset and can re-identify
Journalist scenario	Organizational discredit knowing there is at least one person that can be re-identified
Marketer scenario	Percentage that can be re-identified
Differential identifiability scenario	Analysis performed on two sets, one containing an individual, and one not

Table 5.3. Potential Impact of De-identified Data. Source: [11].

Potential Impact of De-Identified data	
Identity disclosures	Specific data linked to a specific individual
Attribute disclosures	A piece of confidential information can be attributed to a subject
Inferential disclosures	Information inferred with high confidence from data statistics

Table 5.4. Adversary Skill Levels. Source: [11]

Adversary Skill Level	
General Public	Anyone with access to public information
Expert	A computer scientist skilled in re-identification
Insider	A member of the organization which produced the data
Insider Recipient	A member of the organization which receives de-identified data but has access to other background information
Information Broker	Gathers both de-identified and identifying data, combined into a larger set for exploit
Nosy Neighbor	Friend or family member with access to specific context

Table 5.5. Privacy Risk Harms in De-identification Disclosure. Source: [11].

Privacy Risk Harms in De-Identification Disclosure	
Identity disclosure	Insufficient de-identification
	Re-identification by linking
	Pseudonym reversal
Attribute disclosure	Confidential data release
Inferential disclosure	Group harms

Table 5.6. Data Categories, Examples, and Mitigation Approaches. Source [12]

Data Categories	Example	Approach
PII direct identifiers	Name, SSN, e-mail	Remove where possible Aggregate
Aggregation variables	Birthdates, ages, locations	Generalizations Replace with ranges
Demographic indirect identifiers	Sex, ethnicity, occupations	Remove where possible Aggregate
Outlier variables	Medical procedures performed Distinct deformities	Assess risk and remove if necessary
Structured data variables	Vital signs, lab tests, and results	Perform re-identification risk analysis
Freeform text	Physician notes, referral letters	Omit PII from freeform text Natural language processing
Non-parsable voice	Voice recordings	Remove
Image data	X-rays, scans	Omit where possible

5.3 Methodology

The methodology proposed in this Chapter is to apply the fundamentals of the RMF and PRMF with the circumstances of the RDC and our data sharing model.

5.3.1 Categorize

The types of PI in the RDC are quite vast. Since disk images are derived from real people, PII types are not strictly confined to one specific area. Disk images can contain financial, medical, professional information, etc., in unstructured formats. Our model however, precludes us from performing de-identification on the RDC drives. Instead we focus the results of the algorithm run on the RDC and will derive potential PII information types from the output. Either by communication with the extramural researcher or testing arbitrary code on the RDC, each scenario will categorize and identify PII types. Ways in which we might determine and categorize PII on scenarios.

- Determine the algorithm's purpose?
- How much access has the algorithm been granted and what files types are accessible?
- What PII types were observed, if any?
- What is the format of the algorithm's output?
- Are there any problematic data actions?
- Manually review the output and view if structured, semi-structured or unstructured format. Is the output in binary?

5.3.2 Controls

Once we have discovered what PII types are in our system and the categories of data that we are working with, we select the appropriate controls to mitigate the potential risk. The following are steps to identify what controls can be used.

- What application can successfully use to translate the data?
- What tools can we use to de-identify the algorithm data?
- What data sharing model can be used to protect PII?
- What tools can we use to attempt to re-identify algorithm data?
- What access restrictions can be applied to secure the data?

- Use PII forensic tools like bulk_extractor or perhaps other free form text processing tools or scrubbers.

Once controls are selected, we must implement them. Here we test the algorithms on a small sample of the RDC to observe behavior. Anonymized results and in relation to our taxonomy of risk scenarios, make a risk assessment before release to extramural researcher

5.3.3 Assess

In the assess step, the process identifies risks, assets, value, and harm in regards to the data subjects and organization. Assessments start off with defining what parameters are via the framing of objectives from our first step. We will identify threats, events, or vulnerabilities and can then assess impact and how these factors affect the data subjects. In regards to privacy assessment we will also identify specific risks relating to privacy of the individual.

- Assess the possible threat sources.
- Assess characteristics of the threat in regard to capability and intent.
- What factors mitigate these threats?
- Evaluate the likelihood that the threats will be initiated.
- What vulnerabilities make these threats more likely?
- What is the likelihood of a threat succeeding?
- What are the impacts of PII release?
- Assess risk, based on impact and overall likelihood using Table 5.7.

Table 5.7. Risk Assessment Scale. Source [2].

Likelihood (threat causes impact)	Level of Impact				
	Very Low	Low	Moderate	High	Very High
Very High	Very Low	Low	Moderate	High	Very High
High	Very Low	Low	Moderate	High	Very High
Moderate	Very Low	Low	Moderate	Moderate	High
Low	Very Low	Low	Low	Low	Moderate
Very Low	Very Low	Very Low	Very Low	Low	Low

5.3.4 Authorize

Authorization, or determination to release, can be made once the controls that were implemented are assessed. If appropriate controls are implemented to reduce the likelihood and

impact of a threat event, the risk objectives of the organization are met, the organization can make a determination to release the data.

- Did the risk assessment fall within the boundaries of the goals and objectives set by the organization?
- Were the organizational goals met?
- Determine if the risk is acceptable.

5.3.5 Monitor

In the monitor step, evaluation is done to test if controls or responses are effective. Continual monitoring of our system allows the researchers to make adjustments and to adapt to various changes or include new remediation (i.e., new identifiers) into our process. It is important to note that for future research every scenario conduct their own assessment and determine what types of PI they have observed and dealt with or de-identified. Whether using pseudonyms or statistical disclosure limitation, all these methods should be documented to keep track of what has been released and modified. A well-documented system also assists future researchers to learn from previous experiments and monitor re-identification attacks. As the "data rich network" evolves pseudonyms become more prone to re-identification attacks [11]. Especially in the case with pseudonyms where overtime they can be reversed. Here we confirm that the algorithm is running on the RDC as intended. After disclosure we will review the work of the qualified researcher, and any findings or published material regarding our dataset to check results.

5.4 Organizational Requirements

The scope of PII problems can become unmanageable due to various data modalities highlighted in previous sections. Therefore, we placed the following requirements on our scenarios with requests to run on the RDC.

- Conditions of use of the RDC meets the purpose and goals of advancing the state-of-the-art of digital forensics
- Objectives are derived from the standards of ethical conduct and research established by the IRB. The DOD affirms these standards.

- Organizational and research objectives state that the welfare of data subjects are a primary concern and PI we consider above minimal risk, will not be disclosed.
- The PI data is not being used for the sole purpose of identifying data subjects
- Agree that results of algorithm and analysis will be checked for PI and go through de-identification process.
- Due to our resources we are not able to de-identify images, audio, or video, sources of information.
- Data sharing model implemented is a hybrid between enclave data release model and interactive query interface model(limited to qualified researchers).
  - Export of data is monitored by a data controller like that of an enclave, but queries are run on the original RDC data set (like that of the interactive query interface).
  - The results from original data set are then reviewed and go through a process of de-identification(depending on the modality).
  - Data controller then reviews and checks de-identified data set for potential PII leaks. If satisfied and deemed low risk, data controller will disclose to requesting party.

Researchers looking to extract data clearly for PII and after the anonymization process find results to be useless would not benefit from running their algorithm on the RDC nor do our policies allow for it therefore it is with the explicit understanding that the extramural researcher understand that all direct identifiers, in addition to some quasi-identifiers will be sanitized.

Despite the reputation and body of work of the requesting researcher a security evaluation on the algorithm or program being run on the RDC should be done. This is to avoid any unintended information flows that the requesting researcher could be maliciously or unintentionally trying to extract.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 6:
SCENARIO ASSESSMENT

Chapter 5 presented a sample RMF, built to be applicable to the need to de-identify PII and researcher requests to access NPS's RDC. Chapter 6 applies that RMF, step-by-step, to a real world scenario.

6.1 Sifting Collector Scenario

6.1.1 Background

Jonathan Grier, a digital forensic security consultant, in collaboration with Golden Richard III from the University of New Orleans, developed a new evidence acquisition approach called "sifting collectors." Sifting collectors attempt to address the "volume challenge"[18] by hybridizing disk imaging and live memory acquisition methods with the goal of extracting only relevant data. Thus, sifting collectors can recognize and identify relevant regions of the disk and image only those regions without losing important artifacts or risking PII disclosure from other areas of the disk [77].

After reviewing Grier's paper on sifting collectors and correspondence, NPS RDC researchers were able to determine his motivations for RDC access. Because the sifting collector relies on identifying relevant regions of memory and performs selective acquisition, researchers understood that the RDC's data could provide a wealth of information in helping Grier identify such regions, improving the tool's accuracy and efficacy on many real world drives. Based on the core requirements described in Chapter 6, researchers understood that PII was not one of their research goals, and they had no reservations about the de-identification of PI. Grier was also willing to provide their source code and additional assistance in helping to test their program on the RDC, so researchers determined that Grier and Richards were qualified requesters and that their research would benefit the digital forensic field. The researchers had completed their assessment.

[18]Coined by Doug Laney, the volume challenge attempts to address one of the four Vs that characterize big data; the Vs are described in Section 2.2.

6.1.2 Methodology

The following steps were taken to analyze risk:

- Identify level of access classification needed as complete access
- Classify sifting collector as arbitrary code with source code provided
- Assess algorithm by manually testing on small set of simulated data. Perform security evaluation on input algorithm while concurrently observing the characteristics of the output.
- Use bulk-extractor to extract any potential PII on results. None were found.
- Observe and classify output data and identifiers, determine grain-set file as binary and log files as unstructured text file. No direct identifiers relating to human data subjects presented. Quasi-identifiers were potentially indicated, specifically, path names of disk images that identified the country of origin from which the disk drive was purchased.
- Apply algorithm throughout the data corpus using multiprocessing script, then review output created.
- Run bulk_extractor again on all scanners to identify any potential PII. None were found.
- Manually sample and review results for potential PII exposure. If found, document and categorize that identifier. Try to adjust bulk_extractor scanner or script to account for that information type. None were found in this scenario.
- Researcher was not given actual disk images, only diagnostic and grain-set files. The researcher was interested in file system profiles, in order to apply the RDC to compare probabilistic areas of memory that might have forensic relevance. The output, however, was not a disk image but a grain-set, information about the structure of the disk image. As actual data was not transferred to the researcher, privacy was fully maintained.

6.1.3 Testing and Output

Researchers utilized the sifting collector's evaluation mode, since the RDC is a organized collection of preexisting E01 raw images. The sifting collector was initially run on simulated data found in the RDC. After reviewing the results, researchers then moved forward with running the program in bulk using a python script and its multiprocessing module. A

successful run of the sifting collector yielded the following outputs for each image:

- *-sifted.E01
- ground-truth.gst
- sifted.gst
- diagnostic log file

The program then compressed three of the outputs, leaving out the sifted images naming convention *-diagnostic.zip. Due to storage constraints of the RDC, our script incorporated the deletion of sifted images but retained all files in the zip file.

6.1.4 De-Identification measures

After the initial and full run of Grier's sifting collector on the RDC's 3,098 disk images, researchers received 1,319 results from the algorithm. Each disk image on the RDC has a corresponding cryptographic md5 ID hashes, which is a digital fingerprint of the disk image. Out of the 1,319 results, researchers could only locate 909 of these hashes, perhaps due to errors during disk imaging process. After reviewing diagnostic logs, researchers came to the determination to mask the path names to RDC images. This was done, primarily, to prevent extramural researchers from knowing the layout of our file system and, also to remove the country of origin which is located in the path name. The choice of transforming abbreviations for the country names using hashes is a form of pseudonymization which preserves the relationship for later use by the data controller, but leaves the system open for future re-identification. After running the results on bulk_extractor to check for any attributes or PII, researchers then proceeded to extract all the path names and placed them into a file. After retrieving 909 md5 hashes, researchers then wrote a script to write over the pathname in Grier's diagnostic.log files.

6.1.5 Analysis

After receiving and installing the sifting collector, using an unprivileged account, researchers ran the program in evaluation mode[19] on a Non-U.S. (NUS) directory of the RDC using server @domex.nps.edu on Linux.

[19]evaluation mode on Grier's program ran on disk images

It is worth mentioning that Grier's sifting collector currently only works on New Technology File System (NTFS)[20] file formats. Thus, the sifting collector did not collect disk image information on those that have been reformatted, damaged, or tampered with on a disk level to hide data.

Researchers also observed, while working on this scenario, that performing a security evaluation on proprietary source code was laborious. Although trying to avoid PII disclosure through an exploit or vulnerability in his program, this process of evaluating code would not be feasible within the de-identification model, and researchers questioned if the model would take any arbitrary algorithms.

The diagnostic log file tracks the programs status during the processing of an image. The contents tracks the grains and size of the disk in addition to the date and time the process ran on the RDC. Other information includes number of partitions, file system format of those partitions, unallocated or allocated regions, number of nodes within the file system, and the percentage and time the sifting collector took to process the image. What was concerning, however, was that the log file captured the path and names of the images, which both divulges the disk image's country of origin and reveals the layout of the RDC database. Although country location alone is not considered a sensitive personal identifier it would take minimal effort to anonymize such data, and anonymization would reduce the possibility of re-identification via quasi-identifiers and linking of database information through results and analysis of other research results.

Since the objective of the sifting collector is to run on disk images and analyze, sector by sector, what areas were forensically relevant and then copy those regions, researchers allow Grier's tool to access all types of data which is a flag for a high risk classifier. However, the sifting collector is not interested in the content of the data but rather its relevance and potential as evidence. As Grier explains in his request, they seek "grains containing any data associated with forensically relevant file... or at least one forensically relevant disk sector" [77]. If a sector is imaged, then a 1 is written as output for that sector in a grain-set (GST) file, 0 if not copied. This is a low level abstraction that provides no file content information.

[20]NTFS developed by Microsoft as a high performance file system built on top of the FAT file system with more permission features, journaling, and uses a special file called a Master File Table that handles metadata and allocates spaces for files more efficiently [78]

The *.gst files produced are grain-set files, which are binary representations of what the sifting collector copied. The grain-set file produced a binary output which makes it hard for human readability and manual processing of data. Because binary formats are considered high risk in output classification, we would have to perform a thorough analysis of the source code.

6.2 Determination for Disclosure

To determine risk of disclosure for this scenario, we will consider the likelihood of a threat event based on the relevant threats and RDC vulnerabilities, and how they relate to the impact or harm of an data subject if the output was released. The risk assessment table developed by NIST, shown in Table 5.7, is used to determine risk as a function of likelihood and impact.

6.2.1 Likelihood

Threats and vulnerabilities affect how likely a threat event is to occur. We noted the following attributes from the scenario, which we then assessed to see if each would increase or decrease the likelihood of the threat event's occurrence.

- High adversary skill: Increase
- Trusted adversary reputation: Decrease
- Full access to disk image : Increase
- No availability of external data links: Decrease
- Algorithm source code known: Increase
- De-identification by pseudonymization: Decreases
- Unstructured text output: Increase
- Re-identification: Decrease

Taking all the threats and vulnerabilities into account, we assess that the likelihood of data compromise is very low. The most dominant factor is that the actual adversary in this case is a trusted agent (extramural researcher).

Table 6.1. Scenario Risk Assessment. Source [2].

Likelihood (threat causes impact)	Level of Impact				
	Very Low	Low	Moderate	High	Very High
Very High	Very Low	Low	Moderate	High	Very High
High	Very Low	Low	Moderate	High	Very High
Moderate	Very Low	Low	Moderate	Moderate	High
Low	Very Low	Low	Low	Low	Moderate
Very Low	Very Low	Very Low	Very Low	Low	Low

6.2.2 Impact

As discussed earlier, the only identifying information that could be released is the country of origin. As a quasi-identifier, the impact is assessed as low. It would take more than just this one attribute to identify an individual.

6.2.3 Risk Determination

After assessing the likelihood that the threat event would occur as unlikely and that the impact of the release of the data would be low, we assessed an overall risk of very low as shown in Table 6.1. With a very low risk level, we have high confidence in granting Grier's request.

CHAPTER 7:
CONCLUSION AND FUTURE WORK

7.1 Conclusion

The goal of this research was to share the RDC while maintaining privacy of data subjects while also providing the maximum availability of useful data.

This thesis discussed a method of risk assessment as applied to scenarios that contain PII. We highlighted the definitions, legal and ethical concerns, and technical aspects of allowing extramural researchers access to the RDC. A taxonomy and methodology for approaching this issue was developed and we were able to answer the following questions.

What are the risks, and what is considered acceptable risk of disclosure? Risks were evaluated based on the probability of the threat and the impact of disclosure. Various threats were outlined in Chapter 5 as well as different impacts. Each scenario will be different, but the methodology in Chapter 5 can be used by NPS again to evaluate what level of risk is involved when working with extramural researchers.

How can we allow extramural researchers access to the RDC and institutions without significant risk to human subject privacy? We are able to do this by implementing various de-identification tools and security control measures. After assessing the threat likelihoods and impact, we identified the scenario as low risk. Our background research revealed that HIPAA's Safe Harbor de-identification standard was the best current method for removing PI, direct and quasi-identifiers, from our results. Although we did not find any PI identifiers eighteen Safe Harbor identifiers, we saw fit to remove the abbreviated country names on the RDC pathnames. Masking of pathnames would not reduce the utility of the data set but obscured a potential link between the data set and human data subject.

Due to the heterogeneity of data, can we effectively build a criteria for algorithms and how restrictive must the criteria be to protect human subject confidentiality? We developed a set of restrictions to bound the problem so that our methodology could be applied effectively. The restrictions placed on extramural researchers included the following.

- Working inside IRB guidelines, which was achieving minimal risk
- PII not used solely to identify individuals
- Agree that algorithms will be checked and vetted.
- Only text-based outputs could be used
- Requests be fulfilled using our hybrid data sharing mode.

Can we successfully de-identify PII output generated by vetted algorithms provided by external researchers and safely disclose the results? Tools such as bulk_extractor and our own scripts allowed us to verify that there was no PII in the output that was provided to the extramural researchers. As the country of origin could be gleaned from the path names, it was a relatively easy task to develop a script to hide the country names.

At what point do results lose their utility when too much PII is removed? We discussed in this thesis that different levels of access can severely hinder the objectives of the research. We also saw that aggressive de-identification could alter the data too much and lead to less accurate results. A dialogue must be established by NPS and extramural researchers to assess on a case by case basis, the level of access and de-identification required. Also NPS should also adopt a DUA along with continued monitoring with IRB to prevent third parties from re-identification and save procedures to control data release.

We can improve on future research by establishing a procedure for continuous monitoring and logging of the PII encountered in various scenarios and how that PII was processed. We discussed that keeping a record of this allows future researchers to go back into previous work and improve on what had been done.

7.2 Future Work

The framework forms the basis for future work in de-identification and data release procedures. When first investigating de-identification our goal was to find scenarios and use digital forensic tools to try and automate a de-identification process. After our first scenario and researching into another request, we found that de-identification was a problem with a huge scope, which not only had mathematical complexities but dealt with the legal and ethical issues of privacy, which is hard to define in itself. With the task of de-identification of human data subjects, it is required to understand the mechanisms, standards, and proce-

dures to release de-identified information responsibly. However, further work needs to be done to understand how PI is being used.

A natural follow on to our project would be to: Build a synthetic dataset of the RDC and use the Synthetic Validation Model to help extramural researchers develop and perfect tools, then offer SDL results with tests run on the original data set, for validation.

Much of our restrictions were confined to text-based structured and unstructured files. As one of our restrictions was to use text based files, we must look into other types of media and develop tools to identify PII on images, video, and other mixed media. More tools need to be developed and identified so we can add biometric identifiers to our study.

After creating synthetic or de-identified datasets, we may want to develop a more thorough understanding of re-identification methods to better evaluate the threat likelihood. Exploring re-identification attacks can help researchers better protect datasets. Once these methods are known, we should attempt to apply these re-identification techniques on algorithm outputs before disclosure.

When working with unstructured data sets, language processing must be incorporated into tools to account for free text. Better identification of information types in free file formats will help reduce unintentional leaks in addition to threats posed by an adversary.

On a grander scale, privacy preserving models for information systems is something cybersecurity professionals should be looking into. Cybersecurity is not always individual privacy security but the study of protecting information will benefit all in the security field.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX: Other Definitions and Terminology

A.1 NIST Risk Management Framework

An organization wide security program which focuses on management of organizational risk - risks involving the organization and all individuals within and apart the information system. Risk Management Framework takes a risk-based approach when deciding on security controls and configurations in addition to achieving effective and efficient system while still abiding by laws, ordinances, directives, policies and regulations. Risk Management Framework is a step process that incorporates multiple documents. Below are the six steps and the corresponding publications that address development.

- **Step 1**:Categorize
  - Deals with information systems, the process, storage and transmission of data determined by impact analysis.
  - Refers to FIPS 199 as guidelines for legislative, policy, directive, regulation, standards and organizational directions and the identification of their security requirements.
- **Step 2**: Select
  - Focuses in baseline security controls and categorizes security control baseline using possible risk assessments that the organization has made in reference to their specific conditions.
  - Refers to NIST SP 800-53
- **Step 3**: Implement
  - How to execute implementation of security controls and how to make record of security and how they are used within the organizational information system and environment.
- **Step 4**: Assess
  - Determine if security controls implemented have been done so as planned, that they are working correctly, and they are fulfilling their purpose and meeting security requirements.
  - NIST SP 800-53A Security Control Assessment Procedures

- **Step 5**: Authorize
  - Information system operations are chosen on the basis of determination of risk and takes into account the organizations assets, individuals, other organizations, and the country and determining if choices are of acceptable risk.
  - NIST SP 800-37 Revision 1 provides guidelines on the authorizations of operational information systems.
- **Step 6**: Monitor
  - Assess selected security controls and information system in an ongoing nature to check for effectiveness, and documenting if changes are needed. Also conducting a security impact analysis on any changes made and reporting the state of security.
  - NIST SP 800-37 Revision 1 gives monitoring procedures based on security controls and environment. Also helps determine ongoing risk determinations and help approve authorization to operational status.

A.2 NIST Special Publications and Document Summaries

A.2.1 NIST SP 800-53: Security Privacy Controls for Federal Information System and Organization

A.2.2 NIST SP 800-37: Applying the Risk Management Framework to Federal Information Systems

A.2.3 NIST SP 800-39: Managing Information Security Risk

A.2.4 NIST SP 800-30R1: Guide for Conducting Risk Assessment

A.2.5 NIST SP 800-100: Information Security Handbook a Guide for Managers

A.2.6 NIST SP 800-53: Security and Privacy Controls for Federal Information Systems and Organizations

Table A.1. Definitions of Security Objectives between FIPs 199 and FISMA. Source: [13].

Security Objective	**FISMA Definition**	**FIPs 199 Definition**
Confidentiality	"Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information. . . "	A loss of confidentiality is the unauthorized disclosure of information.
Integrity	"Guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity. . . "	A loss of integrity is the unauthorized modification or destruction of information.
Availability	"Ensuring timely and reliable access to and use of information. . . "	A loss of availability is the disruption of access to or use of information or an information system.

A.3 Examples of Frameworks, Methodology, and Assessments

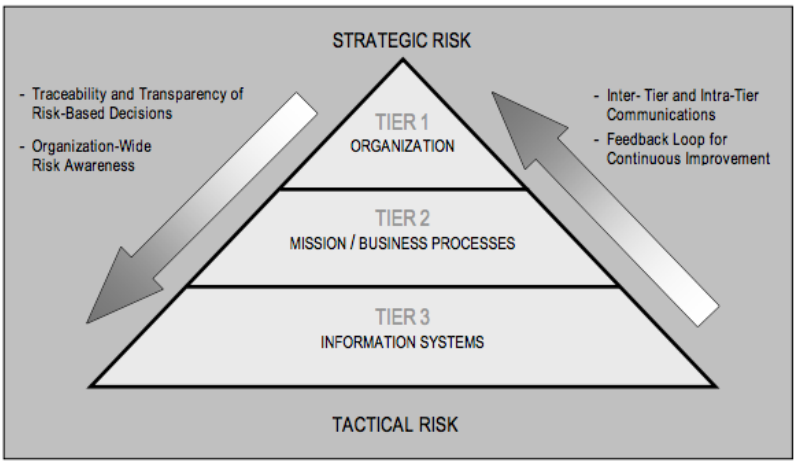

FIGURE 2: MULTITIERED ORGANIZATION-WIDE RISK MANAGEMENT

Figure A.1. Strategic Risk Chart of Risk Management and Assessment as Applied throughout the Tiers of an Organization. Source: [5].

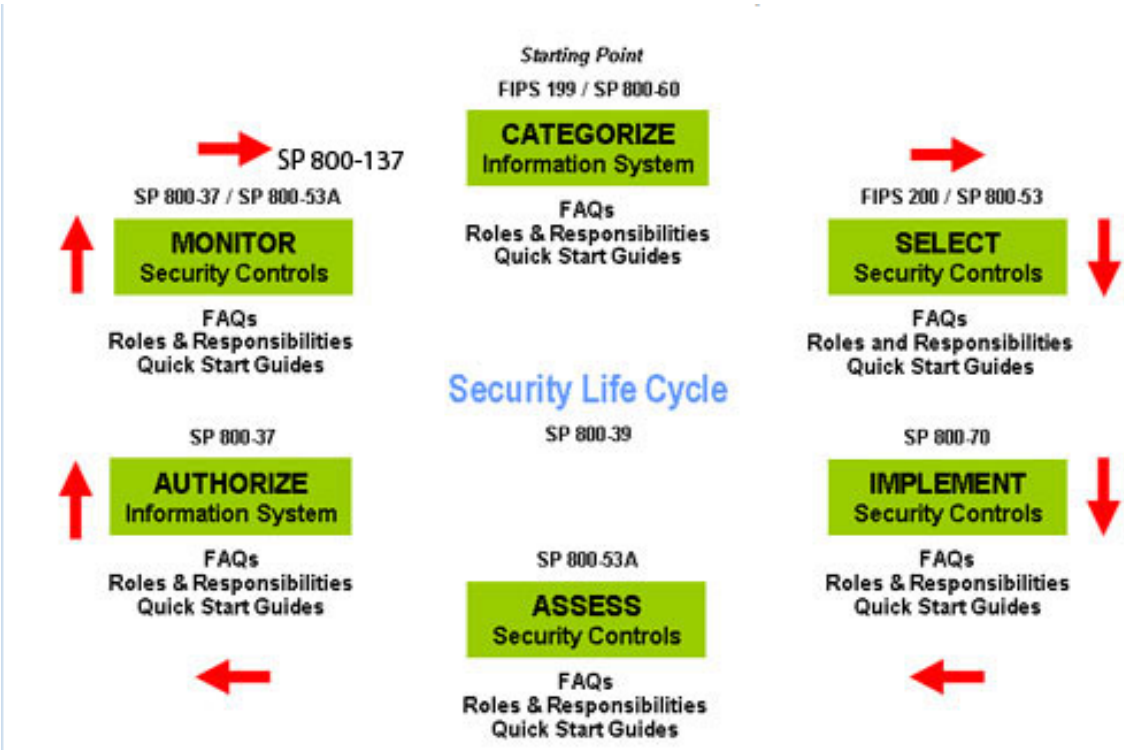

Figure A.2. NIST Risk Management Framework Security Life Cycle Illustrating Six Steps for Risk Management and the NIST SP and Federal Documents that Provide Guidelines. Source: [5].

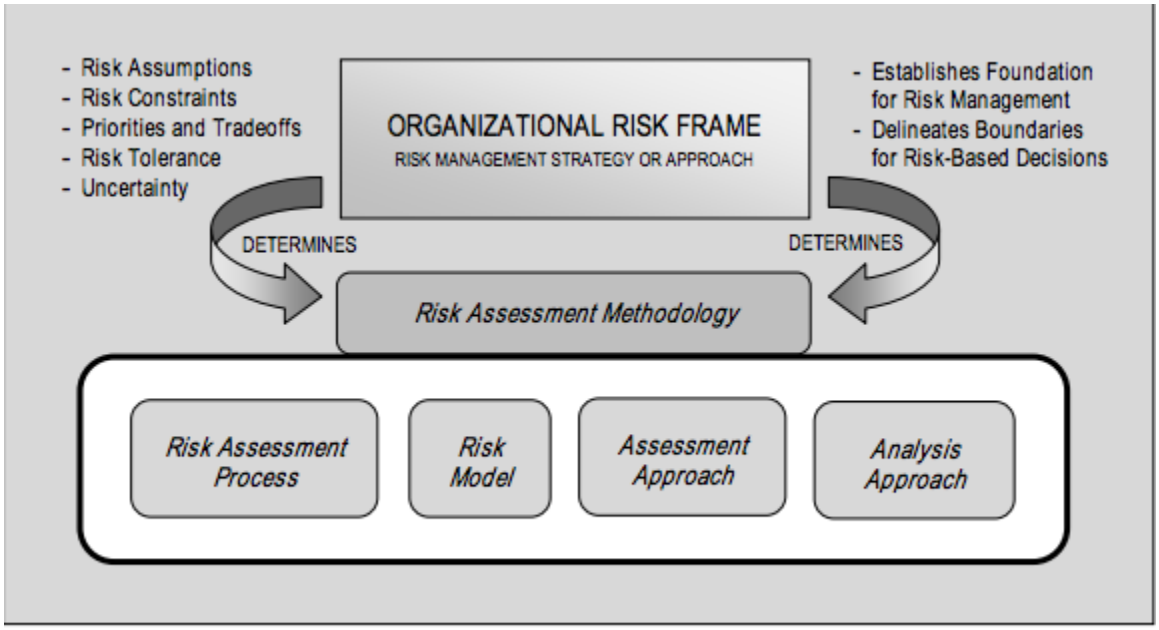

FIGURE 2: RELATIONSHIP AMONG RISK FRAMING COMPONENTS

Figure A.3. NIST Risk Assessment Methodology from Risk Management to Risk Assessment Four Steps. Source: [2].

Table A.2. NIST Example of Threat Taxonomy. Source: [2].

TABLE D-2: TAXONOMY OF THREAT SOURCES

Type of Threat Source	Description	Characteristics

ADVERSARIAL
- Individual
  - Outsider
  - Insider
  - Trusted Insider
  - Privileged Insider
- Group
  - Ad hoc
  - Established
- Organization
  - Competitor
  - Supplier
  - Partner
  - Customer
- Nation-State
	Individuals, groups, organizations, or states that seek to exploit the organization's dependence on cyber resources (i.e., information in electronic form, information and communications technologies, and the communications and information-handling capabilities provided by those technologies).	Capability, Intent, Targeting

ACCIDENTAL - User - Privileged User/Administrator	Erroneous actions taken by individuals in the course of executing their everyday responsibilities.	Range of effects

STRUCTURAL
- Information Technology (IT) Equipment
  - Storage
  - Processing
  - Communications
  - Display
  - Sensor
  - Controller
- Environmental Controls
  - Temperature/Humidity Controls
  - Power Supply
- Software
  - Operating System
  - Networking
  - General-Purpose Application
  - Mission-Specific Application
	Failures of equipment, environmental controls, or software due to aging, resource depletion, or other circumstances which exceed expected operating parameters.	Range of effects

ENVIRONMENTAL
- Natural or man-made disaster
  - Fire
  - Flood/Tsunami
  - Windstorm/Tornado
  - Hurricane
  - Earthquake
  - Bombing
  - Overrun
- Unusual Natural Event (e.g., sunspots)
- Infrastructure Failure/Outage
  - Telecommunications
  - Electrical Power
	Natural disasters and failures of critical infrastructures on which the organization depends, but which are outside the control of the organization.
	Note: Natural and man-made disasters can also be characterized in terms of their severity and/or duration. However, because the threat source and the threat event are strongly identified, severity and duration can be included in the description of the threat event (e.g., Category 5 hurricane causes extensive damage to the facilities housing mission-critical systems, making those systems unavailable for three weeks).
		Range of effects

Table A.3. NIST Example of Threat Assessment Scale. Source: [2].

TABLE D-3: ASSESSMENT SCALE – CHARACTERISTICS OF ADVERSARY CAPABILITY

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	The adversary has a very sophisticated level of expertise, is well-resourced, and can generate opportunities to support multiple successful, continuous, and coordinated attacks.
High	80-95	8	The adversary has a sophisticated level of expertise, with significant resources and opportunities to support multiple successful coordinated attacks.
Moderate	21-79	5	The adversary has moderate resources, expertise, and opportunities to support multiple successful attacks.
Low	5-20	2	The adversary has limited resources, expertise, and opportunities to support a successful attack.
Very Low	0-4	0	The adversary has very limited resources, expertise, and opportunities to support a successful attack.

TABLE D-4: ASSESSMENT SCALE – CHARACTERISTICS OF ADVERSARY INTENT

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	The adversary seeks to undermine, severely impede, or destroy a core mission or business function, program, or enterprise by exploiting a presence in the organization's information systems or infrastructure. The adversary is concerned about disclosure of tradecraft only to the extent that it would impede its ability to complete stated goals.
High	80-95	8	The adversary seeks to undermine/impede critical aspects of a core mission or business function, program, or enterprise, or place itself in a position to do so in the future, by maintaining a presence in the organization's information systems or infrastructure. The adversary is very concerned about minimizing attack detection/disclosure of tradecraft, particularly while preparing for future attacks.
Moderate	21-79	5	The adversary seeks to obtain or modify specific critical or sensitive information or usurp/disrupt the organization's cyber resources by establishing a foothold in the organization's information systems or infrastructure. The adversary is concerned about minimizing attack detection/disclosure of tradecraft, particularly when carrying out attacks over long time periods. The adversary is willing to impede aspects of the organization's missions/business functions to achieve these ends.
Low	5-20	2	The adversary actively seeks to obtain critical or sensitive information or to usurp/disrupt the organization's cyber resources, and does so without concern about attack detection/disclosure of tradecraft.
Very Low	0-4	0	The adversary seeks to usurp, disrupt, or deface the organization's cyber resources, and does so without concern about attack detection/disclosure of tradecraft.

TABLE D-5: ASSESSMENT SCALE – CHARACTERISTICS OF ADVERSARY TARGETING

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	The adversary analyzes information obtained via reconnaissance and attacks to target persistently a specific organization, enterprise, program, mission or business function, focusing on specific high-value or mission-critical information, resources, supply flows, or functions; specific employees or positions; supporting infrastructure providers/suppliers; or partnering organizations.
High	80-95	8	The adversary analyzes information obtained via reconnaissance to target persistently a specific organization, enterprise, program, mission or business function, focusing on specific high-value or mission-critical information, resources, supply flows, or functions, specific employees supporting those functions, or key positions.
Moderate	21-79	5	The adversary analyzes publicly available information to target persistently specific high-value organizations (and key positions, such as Chief Information Officer), programs, or information.
Low	5-20	2	The adversary uses publicly available information to target a class of high-value organizations or information, and seeks targets of opportunity within that class.
Very Low	0-4	0	The adversary may or may not target any specific organizations or classes of organizations.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

[1] S. Al-Fedaghi and A. Al-Azmi, "Experimentation with personal identifiable information," *Intelligent Information Management*, vol. 4, no. Issue 4, pp. 123–133, July 2012.

[2] *Guide for Conducting Risk Assessments*, NIST SP 800-30, 2012.

[3] M. Garcia *et al.*, "Privacy risk management for federal information systems," NIST, Washington DC, Tech. Rep. IR 8062 Draft, May 2015.

[4] L. Sweeney. Data Privacy Lab. [Online]. Available: http://dataprivacylab.org/projects/identifiability/paper1.pdf

[5] *Managing Information Security Risk - Organization, Mission, and Information System View.*, NIST SP 800-39, 2012.

[6] Department of Navy (DON) Chief Information Officer (CIO) Privacy Team. (2011, July). What is Personally Identifiable Information? [Online]. Available: http://www.doncio.navy.mil/ContentView.aspx?id=2428. Accessed August 15, 2016.

[7] *DOD Privacy Program*, DOD Directive 5400.11-R, Deputy Secretary of Defense, Washington DC, 2014, pp. 1–17.

[8] *Guide for Protecting the Confidentiality of Personally Identifiable Information*, National Institute of Standards and Technology (NIST) SP 800-122, 2010.

[9] *Standards for Security Categorization of Federal Information and Information Systems*, Federal Information Processing Standards (FIPS) PUB 199, 2004.

[10] U.S. Department of Health and H. Services. (2007, Feb.). HIPAA Privacy Rule. [Online]. Available: https://privacyruleandresearch.nih.gov/pr_08.asp. Accessed September 9, 2015.

[11] S. Garfinkel, "De-identification of personal information," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. IR-8053, Oct. 2015.

[12] Electronic Transmission of Personally Identifiable Information. (2014). Integrating Healthcare Enterprise. Unknown. [Online].

[13] *Volume 1: Guide for Mapping Types of Information and Information Systems to Security Categories*, NIST SP 800-60v1r1, 2008.

[14] L. Sweeney and M. Crosas, "An open science platform for the next generation of data," *CoRR*, no. 1506.05632, 2015. Available: http://datatags.org/files/datatags/files/sweeneycrosas1.pdf

[15] *The Belmont report: Office of the Secretary Ethical Principles and Guidelines for the Protection of Human Subjects of Research The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*, Belmont Report, 1979.

[16] S. Garfinkel *et al.*, "Bringing science to digital forensics with standardized forensic corpora," *Digital Investigation - 12th Annual Digital Forensic Research Workshop (DFRWS)*, vol. 6, pp. S2–S11, 2009.

[17] D. Kahneman and A. Tverskey, *Choices, Values and Frames*, 1st ed. Cambridge, UK: Cambridge University Press, 2000.

[18] "About the department of defense," www.defense.gov/About-DoD, Department of Defense (DOD), 2015.

[19] U.S. Department of Defense (DOD). (2007). Principles of Information. [Online]. Available: http://www.defense.gov/Resources/Principles-of-Information. Accessed August 20, 2016.

[20] *DODI Reduction of Social Security Number (SSN) Use Within DOD*, DOD Instruction 1000.30, Under Secretary of Defense for Personnel and Readiness, Washington DC, 2012, pp. 1–27.

[21] M. Pollitt, "A history of digital forensics," in *International Federation for Information Processing (IFIP) International Conference on Digital Forensics*, Orlando, FL, 2010, pp. 3–15.

[22] National Institute of Justice (NIJ). (2016, June). Forensic Sciences. [Online]. Available: http://www.nij.gov/topics/forensics/pages/welcome.aspx. Accessed July 20, 2016.

[23] *Guide to Integrating Forensic Techniques into Incident Response*, NIST SP 800-86, 2006.

[24] J. Brunty. (2011, Mar. 2). Validation of Forensic Tools and Software: A Quick Guide for the Digital Forensic Examiner. *Forensic Magazine*. [Online]. Available: http://www.forensicmag.com/article/2011/03/validation-forensic-tools-and-software-quick-guide-digital-forensic-examiner

[25] *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*, ISO 3534-1, Std., 2006.

[26] *General Requirements for the Competence of testing and calibration laboratories*, International Organization for Standardization (ISO) Std. 17 025:2005(en), 2005.

[27] *Accuracy (Trueness and Precision) of Measurement Methods and Results*, International Organization for Standardization (ISO) Std. 5725:1994(en), 1994.

[28] National Institute of Standards and Technology (NIST). (2001, Nov.). General Test Methodology for Computer Forensic Tools. [Online]. Available: http://www.cftt.nist.gov/Test20Methodology207.doc. Accessed July 20, 2016.

[29] S. Garfinkel, "Digital forensics," *American Scientist*, vol. 11, pp. 370–377, 2013.

[30] S. Madden, "From databases to big data," *IEEE Internet Computing*, vol. 3, pp. 4–6, 2012.

[31] The Four V's of Big Data. (2013). IBM. http://www.ibmbigdatahub.com/infographic/four-vs-big-data. Accessed June 1, 2016.

[32] C. Walter. Kryder's law. Scientific American. Available: http://www.scientificamerican.com/article/kryders-law/. Accessed August 5, 2016.

[33] B. S. of Information. (2016). What is Data Science?: A New Field Emerges. [Online]. Available: https://datascience.berkeley.edu/about/what-is-data-science/. Accessed July 20, 2016.

[34] I. A. of Privacy Professionals. Glossary of Privacy Terms. International Association of Privacy Professionals website. Available: https://iapp.org/media/pdf/resource_center/IAP,note=

[35] E. A. Q. Adriano, "The natural person, legal entity or juridical person and juridical personality," *Penn State Journal of Laws and International Affairs*, vol. 4, pp. 363–391, dec 2015.

[36] *Health Informatics - Pseudonymization*, International Organization for Standardization (ISO) Std. 25 237, 2008.

[37] *44*, United States Code (USC) Std. Sec. 3541, 2002.

[38] *44*, USC Std. Sec 3542, 2011.

[39] *Information technology — Security techniques — Information security management systems — Overview and vocabulary*, ISO/IEC, 2016.

[40] *Risk Management- Principles and Guidelines*, International Organization for Standardization (ISO) Std. 31 000, 2009.

[41] T. Palys and J. Lowman, "Going boldly where no one has gone before how confidentiality risk aversion is killing research on sensitive topics," *Journal of Academic Ethics*, vol. 8, no. Issue 4, pp. 265–284, 2010.

[42] *Guide for Applying the Risk Management Framework to Federal Information Systems*, NIST SP 800-37, 2010.

[43] Information and P. C. of Ontario. (2016). De-identification Guidelines for structured data. Information and Privacy Commissioner of Ontario's website. Available: https://www.ipc.on.ca/images/Resources/Deidentification-Guidelines-for-Structured-Data.pdf. Accessed July 9, 2016.

[44] *Risk Management framework (RMF) for DOD Information Technology*, DODI 8510.01, 2014.

[45] "Introduction to ethical hacking," Class lecture slides in CS3695, Department of Computer Science, Naval Postgraduate School by Scott Cote, Summer 2016.

[46] *Human Research Protection Program*, Department of Navy (DON) Secretary of the Navy Std. SECNAV Instruction 3900.39D, 2002.

[47] *Protection of Human Subjects and Adherence to Ethical Standards in DOD-Supported Research*, Department of Defense (DOD) Std. 3216.02, 2011.

[48] *De-Identifying Government Datasets*, NIST SP 800-188 Draft, 2016.

[49] P. Ohm, "Broken promises of privacy: Responding to surprising failure of anonymization," *University of California Los Angeles (UCLA) Law Review.*, vol. 57, pp. 1701–2010, Aug. 2009.

[50] S. Garfinkel and A. Shelat, "Remembrance of data passed: A study of disk sanitization practices," 2003.

[51] HHS.gov. (2016, Aug.). Health Information Privacy. [Online]. Available: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#protected. Accessed September 17, 2016.

[52] HHS.gov. (2016, Aug.). Covered Entities and Business Associates. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html. Accessed September 17, 2016.

[53] Structured vs. Unstructured Data. (2012). Bright Planet. Available: https://brightplanet.com/2012/06/structured-vs-unstructured-data/. Accessed Sept 17, 2016.

[54] S. Mittal and J. Vetter, "A survey of software techniques for using non-volatile memories for storage and main memory systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, June 2016.

[55] B. Carrier, *File System Forensic Analysis*, 2nd ed. Upper Saddle River, NJ: Addison-Wesley, 2005.

[56] B. Lazorchak. (2015). Digital Forensics and Digital Preservation: An Interview with Kam Woods of BitCurator. The Library of Congress. Available: http://blogs.loc.gov/thesignal/2015/05/digital-forensics-and-digital-preservation-an-interview-with-kam-woods-of-bitcurator-2/

[57] K. Woods, "Automated redaction of private and personal data in collections," in *Proc. in the Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage*, Vancouver, BC, Sep. 2012, pp. 1–17.

[58] S. Garfinkel. (2014). *bulk_extractor*. [Online]. Available: http://digitalcorpora.org/downloads/bulk_extractor/. Accessed Jan. 8, 2016.

[59] MITRE. (2010). *MITRE Identification Scrubber Toolkit (MIST)*. [Online]. Available: http://mist-deid.sourceforge.net/. Accessed August 29, 2016.

[60] L. B. R. Ltd. (2014, Nov.). The Privacy, Data Protection and Cybersecurity Law Review. [Online]. Available: http://www.sidley.com/~/media/files/publications/2014/11/the-privacy-data-protection-and-cybersecurity-la__/files/united-states/fileattachment/united-states.pdf. Accessed August 3, 2016.

[61] P. Schwartz. (2016, June). Privacy and Security Law: What Korean Companies need to know. [Online]. Available: http://www.paulhastings.com/area/privacy-and-cybersecurity/privacy-and-security-law-what-korean-companies-need-to-know. Accessed August 3, 2016.

[62] Z. A and K. Zetter. (2015, Oct. 8). California Now Has the Nation's Best Digital Privacy Law. *Wired Magazine*. [Online]. Available: https://www.wired.com/2015/10/california-now-nations-best-digital-privacy-law/

[63] *United States Supreme Court CALIFORNIA vs. GREENWOOD, (1988)*, The Supreme Court of the U.S. Std. 486 United States (U.S.) 35, 1988.

[64] Real Data Corpus FAQ. (2012). S. Garfinkel and K. Woods. http://digitalcorpora.org/corpora/disk-images/rdc-faq.

[65] *Human Subject Regulations Decisions Charts*, United States (U.S.) U.S. Department of Health and Human Services (HHS) Std. 45 CFR 46.101(b)(4), 2016.

[66] United States (U.S.) U.S. Department of Health and Human Services (HHS) the Secretary's Advisory Committee on Human Research Problems. Attachment A: Human Subjects Research Implications of "Big Data" Studies. United States (U.S.) U.S. Department of Health and Human Services (HHS) website. Available: http://www.hhs.gov/ohrp/sachrp-committee/recommendations/2015-april-24-attachment-a/index.html. Accessed September 20, 2016.

[67] *Federal Policy for the Protection of Human Subjects*, United States (U.S.) U.S. Department of Health and Human Services (HHS) Std. 45 CFR part 46, 2009.

[68] *DOD Information Security Program: Controlled Unclassified Information* , Department of Defense (DOD) Std. 5200.01 Vol 4, 2012.

[69] *Executive Order 13556*, The White House Std. 13 556, 2010.

[70] U. B. O. of the Chancellor Ethics, "Fair information practice principles (fipps) privacy course," UC Berkeley Website, accessed September 20, 2016. Available: https://ethics.berkeley.edu/privacy/fipps

[71] *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University Data Privacy Working Paper, Rev. 3, 2000.

[72] "K-anonymity and other cluster-based methods," http://webcache.googleusercontent.com/search?q=cache:S1oc98nheBcJ:www.cse.psu.edu/ ads22/course notes/Ruan-k-anon-cluster.ppt+cd=2hl=enct=clnkgl=us, Pennsylvania State University, 2007.

[73] S. Garfinkel, "Testimony and statement for the record of Simson L. Garfinkel, Ph.D." in *Subcommittee on Privacy, Confidentiality Security National Committee on Vital and Health Statistics*, 2016.

[74] Microsoft. (2011, July). Understanding Office Binary File Formats. [Online]. Available: https://msdn.microsoft.com/en-us/library/office/gg615407(v=office.14).aspx. Accessed July 20, 2016.

[75] S. Vadalasetty. (2003, Oct.). Security Concerns in Using Open Source Software for Enterprise Requirements. [Online]. Available: https://www.sans.org/reading-room/whitepapers/awareness/security-concerns-open-source-software-enterprise-requirements-1305. Accessed July 20, 2016.

[76] C. Eagle. (2013, Jul. 17). Ohio Information Security Forum (OISF) 2013 Chris Eagle Reverse Engineering Demystified (a little maybe). [YouTube video]. Available: https://www.youtube.com/watch?v=uhrssX57PkM. Accessed Aug. 10, 2016.

[77] J. Grier and G. Richard, “Rapid forensic imaging of large disks with sifting collectors,” in *Digital Investigation vol. 4 - Proc. Digital Forensic Research Workshop (DFRWS) 2015 US*, Philadelphia, PA, Aug. 2015, pp. S34–S44.

[78] M. H. Ligh *et al.*, *The Art of Memory Forensics*, 1st ed. Indianapolis, IN: John Wiley & Sons Inc, 2014.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California